Statistics

4H/5WP Project Descriptions and Allocations

2020-2021

Project 1: Modelling and forecasting financial time series of different frequencies

Statistics Supervisor: Agnieszka Borowska

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Financial time series exhibit a number of interesting properties which directly affect their modelling and forecasting. Most importantly, they are characterised by time-varying variance, frequently referred to as "volatility", which has a tendency to "cluster", a phenomenon known as "volatility clustering" (see Figure 1).

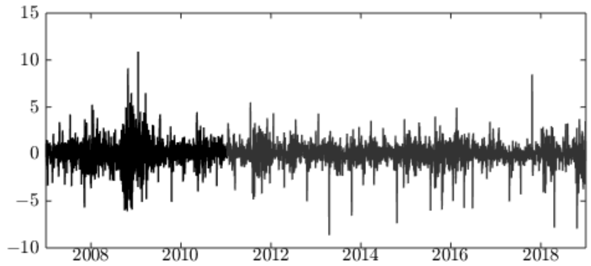


Figure 1 The daily logreturns of the IBM stock from the 4th January 2007 to the 22nd

Different methodologies have been proposed to analyse financial time series, starting from the general Box-Jenkins ARIMA approach, through more specialist GARCH-type models (where GARCH stands for generalized autoregressive conditional heteroskedasticity), up to sophisticated state space models of stochastic volatility.

The aim of this project is to investigate whether different time series frequencies call for different modelling methodologies. You will analyse financial time series of different frequencies (e.g. five minute, daily, monthly) of different financial series (currency futures, i.e. data on futures contracts on exchange rates, stock prices, crude oil futures contract, bond futures).

In particular, the focus will be on forecasting, rather than training-data-fitting (though it is still important to check the latter!). In other words, the models should be evaluated based on their out-of-sample, not in-sample, performance. You will investigate whether the choice of a particular evaluation criterion impacts the model selection process.

The analysis can be carried out from the classical (frequentists) or Bayesian perspective. An interesting extension would be to compare both inference paradigms in terms of their forecasting performance.

References

Bollerslev, T. (1986), “Generalised Autoregressive Conditional Heteroskedasticity”, *Journal of Econometrics*, 51, 307-327.

Engle, R. F. (1982), “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation", Econometrica, 50, 987-1007

Diebold, F. X. and Mariano, R. S. (1995), “Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 253-263.

Kim, S., Shephard, N., and Chib, S. (1998), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models", *The Review of Economic Studies*, 65, 361-393.

Taylor, S. J. (1994), “Modeling Stochastic Volatility: A Review and Comparative Study”, *Mathematical Finance*, 4, 183-204.

Key Questions of Interest

1. Do different time series frequencies (from five minutes to monthly) and the associated time series characteristics require different forecasting methodologies?
2. Are more complex/specialist models profitable, i.e. do they outperform parsimonious/generic models? What is the relationship between model in-sample fit and its out-of-sample performance?
3. How does the choice of a forecast evaluation method affect model selection?
4. Are the generated forecasts superior to the random walk forecasts?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

There are several R packages for time series modelling and forecasting -- the student is supposed to choose most suitable ones for each of the analysed series, get familiar with the corresponding documentation and apply appropriate functions.

In addition, the student will need to perform basic data manipulations, such as converting price series to log return series (if necessary) or dividing datasets into in-sample and out-of-sample parts. Finally, the student will be expected to compare the in-sample and out-of-sample performance from different models.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
Time series
Introduction to R

Desirable:
(Advanced) Bayesian inference

Project 2: Estimation for branching processes

Statistics Supervisor: Alexey Lindo

External Supervisor (if any):

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

Recently a new moment-based parameter estimation techniques for continuous-time, multi-type branching processes has been introduced in [1]. We are going to examine performance of this method via a simulation study.

Literature:
1. Xu, J. et al. (2019) Statistical inference in partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *Annals of Applied Statistics*. 13(4). 2091–2119.

Key Questions of Interest

How well does the loss function estimator performs on simulated data?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 3: Estimation of percolation quantities

Statistics Supervisor: Alexey Lindo

External Supervisor (if any):

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

In this project, via a simulation study, we are going to asses statistical properties of estimators proposed in [2] and apply some of the results described in that paper to some recently proposed models described in [1].

Literature:
1. Duminil-Colpin, H. (2018) Sixty years of percolation. *Proc. Int. Cong. of Math.* **4**. 2847–2874.
2. Meester, R. Steif, J. (1998) Consistent estimation of percolation quantities. *Statistica Neerlandica*. **52**(2). 226–238.

Key Questions of Interest

Can speed of convergence of estimators proposed in [2] be assessed via a simulation study?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 4: Fitting generalised lasso models. A simulation based study

Statistics Supervisor: Alexey Lindo

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

LASSO stands for Least Absolute Shrinkage and Selection Operator and it was originally formulated for linear regression to perform both variable selection and regularisation. The details of the method can be found in~[2]. In this project, we are going to compare performance of several optimisation algorithms for fitting generalised lasso models. Our study will be similar in nature to the one perfromed in~[1].

Literature:
1. Frandi, E. et al. (2016) Fast and scalable lasso via stochastic Frank–Wolfe methods with a convergence guarantee. *Machine Learning*. **104**. 195---221.
2. Hastie, T., Tibshirani, R., Wainwright, M. (2015) ***Statistical learning with sparsity. The Lasso and generalizations***. CRC Press.

Key Questions of Interest

What is the best method of fitting generalised lasso models?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 5: Large-scale hypothesis testing: Empirical null estimation

Statistics Supervisor: Alexey Lindo

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

Multiple hypothesis testing arises when several statistical tests are performed simultaneously on the same dataset and each of these test has a potential of rejecting null hypothesis, i.e. produce a statistical discovery. Traditional methods for multiple comparisons adjustments focus on correcting for modest numbers of comparisons. A different set of techniques have been developed for large-scale multiple testing, in which thousands or even greater numbers of tests are performed. For example, a null distribution is not something one estimates in classical hypothesis testing, while large-scale studies indicate that the theoretical null distribution may fail, see [1] for details. In this project via a simulation based approach we are going to study why the theoretical null may fail.

Literature:

1. Efron, B. (2010) Large-scale inference Empirical Bayes methods for estimation, testing, and prediction. Camridge University Press.

Key Questions of Interest

Why may theoretical null distribution fail in large-scale hypothesis testing?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 6: Mean field asymptotics in high dimensional statistics

Statistics Supervisor: Alexey Lindo

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

Problems that statisticians are dealing nowadays require building complex statistical models with a huge number of parameters. It is common to estimate parameters of models with millions of parameters by iterative optimisation algorithms, see survey [2] for more information. In this project we are going to study high-dimensional limits of statistical estimators and show their similarity to thermodynamic limits of certain statistical meachanics systems. In particualr, via a simulation study we are going to investigate BBAP (Baik, Ben Arous, Peche) phase transition, the principal eigenvector in Principal Component Analysis, see [1] for details.

Literature:
1.Baik, J., Ben Arous, G., Peche, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*. **33**(5). 1643--1697.
2.Montari, A. (2018) Mean field asymptotics in high-dimensional statistics From exact results to efficient algorithms. *Proc. Int. Cong. of Math.* **4**. 2991--3012

Key Questions of Interest

Is there a connection between limits of high-dimensional statistical estimators and thermodynamic limits of disordered statistical mechanics systems.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 7: Random matrices and high dimensional statistics

Statistics Supervisor: Alexey Lindo

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

Consider that data is stored in a matrix *X*, with *n* rows and *p* columns, where *n* represents the number of observations in a data set and *p* is the number of features measured for each observation. Classical statistical theory is mainly concerned with studying the properties of estimators, i.e. functions of data matrix *X*, when *n* is large and *p* is fixed and small. In the last decade, technological advances allowed to collect and store data matrices for which both *n* and p are large. Therefore it is natural to study, at least via simulations, the large *n*, large *p* setting as well. For some recent results in this field, see [1, 2, 3].

1. El Karoui, N. (2018) Random matrices and high-dimensional statistics: Beyond covariance matrices. In: *International Congress of Mathematics*, **3**., pp. 2845–2864.
2. Johnstone, I.M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics.* **29**(2), pp. 295–327.
3. Johnstone, I.M. (2007) High dimensional statistical inference and random matrices. In: *International Congress of Mathematics*, **1**., pp. 307–333.

Key Questions of Interest

What are the properties of the bootstrap as a tool for inference concerning the eigenvalues of a sample covariance matrix computed from an *n* x *p* data matrix *X*, where *n* and *p* are both large?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 8: Statistical analysis of homomorpically encrypted data

Statistics Supervisor: Alexey Lindo

External Supervisor (if any):

Can be adapted to: Single/Combined/MSciWP

Brief Description of Project

The paradigm of computing on encrypted data was introduced by Rivest, et al. in [4], and following that paper the area of cryptography that implements this paradigm is called homomorphic encryption (HE). HE allows to perform a simingly impossible task of processing data without having access to it, see [1] for details. In this project, following [2] and [3], we will study some implications of HE to statical calculations.

Literature:

1. Halevi, S. (2017) Tutorial on Homomorphic Encryption. In ***Tutorials on the foundations of cryptography, dedicated to Oded Goldreich***. Lindell., Y. (editor). Springer.
2. Lu, W., Kawasaki, S., Sakuma, J. (2016) Using fully homomorphic encryption for statistical analysis. https://eprint.iacr.org/2016/1163.pdf. Accessed on the 14th of September 2020.
3. Naehrig, M., Lauter, K.E., Vaikuntanathan, V. (2011) Can homomorphic encryption be practical? In ***Proceedings of the 3rd ACM cloud computing security workshop***. 113--124.
4. Rivest, R., Adleman, L., Dertouzos, M. (1978) On data banks and privacy homomorphisms. In ***Foundations of secure computation***, 169--177.

Key Questions of Interest

Can we perform a large-scale statistical analysis on homomorphically encrypted data?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:
Basic knowledge of R programming language is required to perform simulations.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

All the required statistical methods are going to be covered during the project by the supervisor.

Project 9: Modelling malnutrition among children in Egypt

Statistics Supervisor: Dr Craig Alexander & Dr Amira Elayouty

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Malnutrition, represented in the high prevalence of stunting, increasing rates of underweight, and simultaneously increasing rates of obesity, is a huge burden on Egypt’s economy. Malnutrition is a violation of a child’s right to survival and development and its consequences often remain invisible, until it’s too late. It is therefore important to understand the patterns and determinants of malnutrition in Egypt to inform policy development and program design.

In this project, you will use anthropometric and socio-economic data obtained from the 2014 Egypt Demographic and Health Survey (EDHS2014) for a large sample of Egypt children and their mothers. You will use these data to analyse the situation of nutrition and its determinants among school-age children and adolescents students living in Egypt. It is of interest to investigate the differences in malnutrition prevalence by age, gender, place of residence and socio-economic status factors.

Key Questions of Interest

- Are there any differences in children’s BMI by age, gender, place of residence, socio-economic status?
- How does the mother’s nutritional status impact the child’s nutritional status?
- Are the differences in the BMI distribution the same across the entire distribution?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

While there will be elements of R programming required to fit appropriate models and produce relevant plots, there is no simulation or function development needed for this project.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Linear Models, Generalised Linear Models

Project 10: Characterising features of music using Spotify audio data

Statistics Supervisor: Craig Alexander & Amira Elayouty

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The popularity of music streaming services has surged over the past decade to become the most widely used medium for consuming music. Streaming services provide users access to millions of songs instantly. Such services generate playlists of music for users which can be defined by genre, current popularity and listening habits of the user. Such playlists are generated using statistical methods and multiple sources of data.

In this project, you will use data obtained from Spotify's web API which provides audio features and other relevant information on song tracks. You will use these data to answer several key questions of interest relating to genre and audio features. The data available are vast in terms of features and thus you will also have the opportunity to pose and investigate your own questions of interest. There is also the opportunity for you to obtain your own datasets from Spotify to help answer these questions.

Key Questions of Interest

- How well can we classify music genres using audio features?
- Can we provide reliable predictions of Top 100 chart position for a specific year?
- How does the popularity of songs change over time in terms of their audio features?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will use R to fit classification & regression models.

The student will also be required to code functionality to assess the predictive performance of models and data handling.

There is also the potential to code functionality for web scraping to obtain additional data.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Classification, GLM’s
Desired - Time series analysis, Clustering

Project 11: Modelling child growth in Egypt

Statistics Supervisor: Dr Craig Anderson & Dr Amira Elayouty

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Malnutrition, represented in the high prevalence of stunting, increasing rates of underweight, and simultaneously increasing rates of obesity, is a huge burden on Egypt's economy. Malnutrition is a violation of a child's right to survival and development and its consequences often remain invisible, until it's too late. It is therefore important to monitor children's growth and investigate the patterns and determinants of children's growth in Egypt to inform policy development and program design.

In this project, you will use data from the 2014 Egypt Demographic and Health Survey (EDHS2014). The data include hundreds of variables on the demographic, socio-economic and health features of almost 20,000 surveyed households. This includes cross-sectional measures of child growth and development from across the country. You will use these data to analyze children's growth in Egypt and assess inequalities across the different socio-economic characteristics. It would also be of interest to construct reference growth charts for school-age Egyptian children and adolescents and to compare them with the WHO standards, in order to identify the differences and their public health implications.

Key Questions of Interest

- Are there any differences in children's growth by gender, place of residence and socio-economic status?
- Are those differences the same across the whole age distribution?
- How local reference growth charts may differ from international standards?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate/difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

While there will be elements of R programming required to fit appropriate models (including LM,GLM and quantile regression) and produce relevant plots, there is no simulation or function development needed for this project.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Linear Models, Generalised Linear Models

Project 12: Clustering of zero-inflated data

Statistics Supervisor: Ben Swallow

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Clustering aims to group data such that observations in the same cluster are more similar than those in different clusters. This is commonly done using model-based methods where probability distributions are chosen for cluster assigning, commonly normal distributions or variants of these. These distributions are not particularly successful in cases where there are a large number of exact zeros in the data as these generally exhibit less variation than data without many zeros. This type of data can be common in both ecology (rare species) and biology (levels are too low to measure or errors in the equipment). This project will aim to review literature on methods for clustering zero-inflated data and compare their performance on both simulated and real data exhibiting zero-inflated properties.

Key Questions of Interest

What clustering methods are designed to deal with zero-inflated data?

Which of these methods perform best on simulated data with known clusterings and labels?

How do these perform relative to standard methods e.g. k-means, normal mixture models?

When applied to real data, what do we learn?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

This project will likely involve considerable R programming, although many packages are available for conducting clustering and these will be taken advantage of in the project.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential
Clustering
Classification
(Multivariate methods)

Desirable
Mixture models

Project 13: Generalised linear models for zero-inflated continuous data

Statistics Supervisor: Ben Swallow

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Frequently count data exhibit the phenomenon of 'zero-inflation', where many more zeros are found in the data than would be expected under standard probability distributions (e.g. Poisson, Negative-binomial). A less common case is where non-negative *continuous* data exhibit the same phenomenon. In general, two principal methods have been developed with regards to this problem. The first, named delta models, combine two probability distributions together, the first modelling the proportion of zeros and the second a standard distribution modelling the positive observations. The second approach is to use a family of distributions called the Tweedie distributions, which for specific parameter ranges exhibit zero-inflated behaviour but with a single unified distribution.

The aim of this project is to fit GLM models to compare and contrast the two approaches and see how inference/model fit varies in each case. The algorithms will be run on both simulated data and real datasets to make recommendations on when each method should be used.

Key Questions of Interest

1. Does treating the zeros and positive observations as separate processes provide any additional information? Does it cause any problems?
2. What happens if we fit one model to data simulated from the other approach?
3. Can we conclude which approach is better for a real dataset?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The project will involve using existing R packages to simulate data under both approaches and fit GLMs to these data.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential
GLMs

Desirable
Time series
(Advanced) Bayesian inference

Project 14: Hake abundance in the Mediterranean Sea using a spliced Hurdle Model

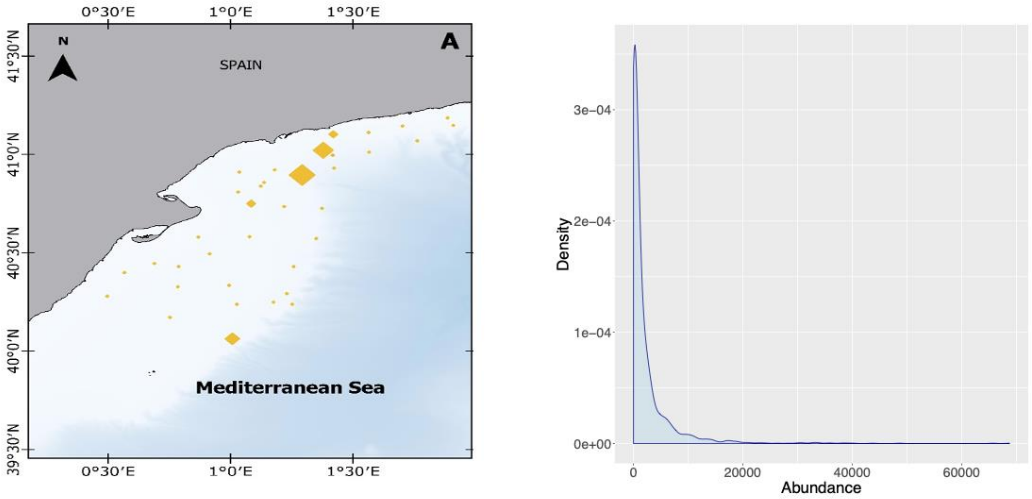
Statistics Supervisor: Daniela Castro-Camilo

Can be adapted to: Single / MSciWP

Brief Description of Project

We are interested in studying the distribution of the European hake in wintertime. The study area is the north-western Mediterranean Sea, between Cape of Salou and Castellon de la Plana (Figure 1a), which includes the area adjacent to the Ebro River Delta. The region covers an extension of 100 km^2 and a depth range from 30 to 350 m, including the continental shelf and upper slope.

Our response variable is hake abundance, which is defined as the number of individuals every 30 minutes of trawling. A simple kernel density estimate for hake abundance is displayed in Figure 1b. We can see a non-negligible proportion of zeros, as well as extreme values. Potential covariates are the capture time, biomass and distance to the coast.



(a)

(b)

We want to develop a model able that adequately captures the whole hake abundance distribution, i.e., the mass point at 0, the bulk of the distribution as well as the right tail. To this end, we start by fitting a Gamma model to the positive hake abundance. This baseline model would likely have a complex additive structure in the mean. Then, we extend this model by implementing the Spliced Gamma-GP model proposed by Castro-Camilo et al. 2019. GP stands for generalised Pareto, which is an asymptotically justified distribution to model values that exceed a certain high threshold, e.g., a value far in the right tail.

The Spliced Gamma-GP model can be viewed as a Gamma model with a tail correction to capture extreme observations adequately. A further extension will consider modelling zero values. We call this modelling approach a **spliced Hurdle model**. A hurdle model is a class of statistical models where a random variable is modelled using two parts; the first is the probability of attaining value 0 and the second one models the probability of the non-zero values. The "spliced" part comes from the fact that we are dividing the distribution of the non-zero values in two (bulk and tail).

A step-by-step analysis plan will be as follows:

1. A basic exploratory analysis.
2. Bulk modelling: Implement a Gamma model for positive hake abundance with an additive structure in the mean.
3. Bulk and tail modelling: Implement the Spliced Gamma-GP model for positive hake abundance.
4. modelling the probability of attaining 0: Extend the Spliced Gamma-GP model to include zero values. This can be carried out by assuming an additive structure in the probability of abundance.

Steps 1-3 constitute the primary analysis and are sufficient for the completion of the project. Step 4 is desirable but not mandatory.

All the models will be fitted using the integrated nested Laplace approximation (INLA), that allows us to numerically approximate posterior distribution of interest. INLA is conveniently implemented in the R-INLA library from the R statistical software.

References

1. Castro-Camilo, D., Huser, R., & Rue, H. (2019). A spliced Gamma-Generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *Journal of Agricultural, Biological and Environmental Statistics*, *24*(3), 517-534. **[For the spliced Gamma-GP model].**
1. **1.** Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395-421. **[For a review on INLA].**

Key Questions of Interest

There are two main applied questions

1. Are there any differences between the baseline Gamma model and the Spliced Gamma-GP model?
2. Is there any significant temporal trend in hake abundance?

Analysis Summary

What level of difficulty do you think the project will have for the typical student? ***Difficult***

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The models should be fitted using the library R-INLA from the R statistical software.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Likely to require

- Bayesian analysis
- Latent Gaussian models (no need to know in advance)
- Generalized additive models
- Integrated nested Laplace approximation (no need to know in advance)

Essential

- Bayesian analysis
- Generalized additive models

Project 15: Landslide Hazard assessment using Hurdle Models

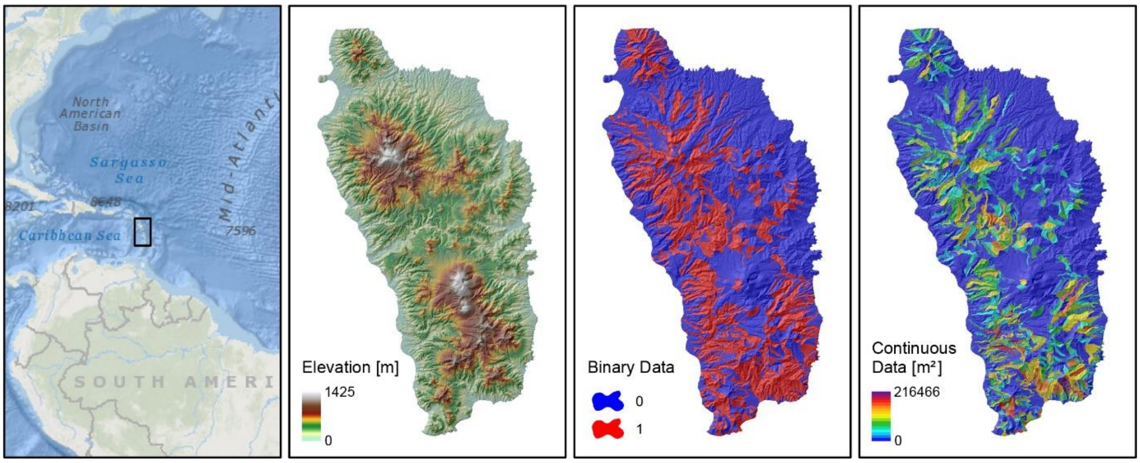
Statistics Supervisor: Daniela Castro-Camilo

External Supervisor (if any): Dr. Luigi Lombardo (Department of Earth Systems Analysis, ITC, University of Twente, The Netherlands)

Can be adapted to: Single / MSciWP

Brief Description of Project

We are interested in studying the joint probability of landslide occurrence and the resulting size. The study area corresponds to the island of Dominica, where Hurricane Maria triggered thousands of landslides on September 18, 2017, reaching category 5 of strength. The island extends for 751 km2, and for this study, it has been partitioned into approximately 4,000 irregular polygons. Each of these polygons has been assigned with two labels. The first one is binary and represents whether a landslide has impacted it or not (occurrence). The second one is continuous and indicates the extent of the landslide inside it (size). Therefore, our response variables are the dichotomous data (see Figure 1c) and the associated continuous landslide size information (see Figure 1d). Potential covariates are continuous (e.g., slope steepness) and categorical (e.g., soil type) in nature, without the use of the actual rainfall because only a weather station recorded the hurricane intensity and duration.



(a) (b) (c) (d)

We want to develop a model able to jointly estimate where a landslide may trigger and how large this may be. This can be achieved using a Hurdle model. A hurdle model is a class of statistical models where a random variable is modelled using two parts; the first is the probability of attaining value 0 and the second one models the probability of the non-zero values. In our case, we can assume a Binomial model for the landslide occurrence, and a Gaussian model for the landslide size.

A step-by-step analysis plan will be as follows:

- A basic exploratory analysis.
- Implement a Binomial model for the landslide occurrence, with an additive structure in the occurrence probability.
- Implement the Gaussian model for landslide sizes, with an additive structure in the mean.
- Produce risk maps to predict the occurrence and sizes of future landslides.

All the models will be fitted using the integrated nested Laplace approximation (INLA), that allows us to numerically approximate posterior distribution of interest. INLA is conveniently implemented in the R-INLA library from the R statistical software.

References

2. Castro-Camilo, D., Lombardo, L., Mai, P. M., Dou, J., & Huser, R. (2017). Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model. *Environmental modelling & software*, *97*, 145-156. **[For the modelling of landslide occurrence.]**
3. Lombardo, L., Opitz, T. and Huser, R., 2018. Point process-based modeling of multiple debris flow landslides using INLA: an application to the 2009 Messina disaster. Stochastic environmental research and risk assessment, 32(7), pp.2179-2198. **[For the use of INLA to model landslide data.]**
4. Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395-421. **[For a review on INLA].**

Key Questions of Interest

1. In the geoscientific community, both responses (landslide occurrence and size) are typically modelled separately. Therefore, a proper joint model will significantly contribute to the field.
2. If we can model the joint probability of landslide occurrence and size, then we can estimate the hazard for future hurricane occurrences. In other words, we would like to estimate the hazard in the Island for a comparable hurricane.

Analysis Summary

What level of difficulty do you think the project will have for the typical student? **Difficult**

Is any Programming/Simulation required? **Yes**

If 'Yes', please specify what this might involve:

The models should be fitted using the library R-INLA from the R statistical software.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Likely to require

- Bayesian analysis
- Latent Gaussian models (no need to know in advance)
- Generalized additive models
- Integrated nested Laplace approximation (no need to know in advance)

Essential

- Bayesian analysis
- Generalized additive models

Project 16: Exploring the language of government and money

Statistics Supervisor: Nema Dean

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Each year the British Chancellor of the Exchequer gives a speech to the House of Commons framing the annual Budget for that year. While some aspects of that speech remain the same over the years ("Mr Deputy Speaker", "I commend this budget to the house", etc.), others may change, reflecting both the different party membership of the chancellor, different economic circumstances, different monetary priorities and different personalities of the speaker! This project will use quantitative text analysis to explore the speeches given over the last 20 years, illustrating changing linguistic trends and the effect of different characteristics of both the speaker and other possible external covariates on the language used to talk about our money.

The student will be responsible for using the quanteda R package (as well as others like tm) to read in textual data and converting this into meaningful quantitative matrices of features. These can then be analysed using various methodologies to allow inference about both the texts and their authors. Possible approaches could include dictionary construction and application, classification and machine learning, scaling models and topic models.

Key Questions of Interest

- Are there major differences in the language used in different budget speeches over time?
- If so, where do these differences lie and are they associated with other covariates?
- Are there certain common topics or keywords across the different speeches?
- Are there differences across different parties or speakers?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

This project will mainly involve coding in R using a variety of text analysis libraries like "quanteda" and "tm". Only real data (not simulations) will be analysed. A good knowledge of using functions and libraries in R is necessary.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

- R coding (essential)
- Linear models (useful)
- Inference (useful)
- Cluster analysis/classification methods (not required but could be useful)

Project 17: Predicting housing tenure in Scotland

Statistics Supervisor: Nema Dean

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The Scottish Household Survey (SHS) is the largest face-to-face survey in Scotland. It is continuous and has been running since 1999. Random repeat sampling is used in order to gain a representative sample of the Scottish public. We want to look at predicting the type of housing tenure a household is likely to have (i.e. social or private renter, owner) based on other household characteristics.

In order to do this we want to use modern machine learning prediction methods like trees and SVMs to achieve a high classification rate but we are also interested in what household variables drive the relationship with tenure. So we are interested in both model performance and generalisability but also inference.

There is a mix of variable types (categorical and numeric) and also missing data issues and recoding to be dealt with in the dataset as well as unequal distribution of tenure types.

Key Questions of Interest

Can we accurately predict the tenure type of a household based on its characteristics?

What characteristics are important in determining household tenure?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

Recoding variables, imputing missing data

Applying classification/machine learning methods

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

- R programming (essential)
- Generalised linear models (essential)
- Multivariate methods (essential)

Project 18: Grouping households in Scotland

Statistics Supervisor: Nema Dean

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The Scottish Household Survey (SHS) is the largest face-to-face survey in Scotland. It is continuous and has been running since 1999. Random repeat sampling is used in order to gain a representative sample of the Scottish public.

We want to look at grouping households in 2018 based on a subset of their characteristics from the SHS to see if particular groups are growing more predominant in recent years in Scotland.

Since there is a mix of variable types (categorical and numeric), the project will start off with classical clustering methods like k-means and hierarchical clustering on separate subsets of the variables (based on type). Later, we will look at designing tailored distance functions for mixed type data as well as looking at modern model-based clustering methods for data of this type.

There are also potential missing data issues and recoding to be dealt with in the dataset.

Key Questions of Interest

Are there different subgroups of households in Scotland?

What characteristics do these subgroups have?

Are there some characteristics more important for separating groups than others?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Recoding variables, imputing missing data

Applying clustering methods

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

- R programming (essential)
- Multivariate methods (essential)

Project 19: Classification Analysis of Acute Respiratory Distress Syndrome.

Statistics Supervisor: Nema Dean

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Acute respiratory distress syndrome (ARDS) is defined as acute hypoxic respiratory failure(PaO2/FiO2<300 mmHg), bilateral chest infiltrates, and the absence of cardiac failure as the primary diagnosis. Treatment aimed at improving survival of this disease is complicated by its extreme heterogeneity. A new treatment thought to improve the disease outcome for patients is Extracorporeal membrane oxygenation (ECMO). Of interest is discovering what biomedical markers both before and after treatment predict the patient's outcome and whether ECMO changes these.

Data are available for 450 patients on biomarkers both before ECMO treatment (marked with a pretext PreECMO, e.g. PreECMO_RR) and for the first day after ECMO treatment (marked with a pretext Day1ECMO, e.g. Day1ECMO_RR).

Key Questions of Interest

Can we use the PreECMO biomedical markers to accurately predict ECMO survival?

Do we need all PreECMO variables or just a subset to make accurate predictions?

What is our expected future performance for these predictions?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

Cleaning data

Plotting data

Applying machine learning methods

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

- R programming (essential)
- Multivariate methods (essential)

Project 20: Spatial modelling of child malnutrition in Egypt

Statistics Supervisor: Dr Amira Elayouty & Dr Craig Anderson

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

In this project you will investigate the causes of malnutrition in children in Egypt. The 2014 Egypt Demographic and Health Survey is a survey of almost 20,000 households across Egypt. The data include hundreds of variables on the demographic, socio-economic and health features of surveyed people. This includes cross-sectional measures of child growth and development from across the country.

There are substantial inequalities in disease risk across Egypt, and we believe that these are driven by health and nutrition inequalities and by deprivation more generally. Therefore we would like to investigate the extent of these inequalities in more detail in order to understand the risks facing children in different parts of Egypt.

The field of disease mapping focuses on showing the extent to which the risk of disease varies across space. We can explore these differences by dividing our study region into a set of non-overlapping subregions (eg counties or local council areas) and computing the overall risk of disease for each subregion in turn. These risks can then be displayed on a map, thus allows us to identify areas which are at high risk, and which may benefit from health interventions.

You will need to identify and fit an appropriate spatial model to these data, and then produce a map of the disease risk across the study region. This will allow you to identify the regions of Egypt with highest and lowest malnutrition.

Key Questions of Interest

1) Is there a spatial pattern in the malnutrition rates across Egypt?
2) Which factors drive the differences in malnutrition between regions?
3) Can individual level effects also be used to explain some of the variability in malnutrition?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will have to use R to fit models such as GLMs, GAMs and spatial/spatio-temporal mixed models.

In most cases, they will be able to learn how to fit these models using existing functions from a variety of packages (eg mgcv, lme4, CARBayes).

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

It would be helpful if the students had some knowledge of the following techniques, but they will should be able to learn the relevant skills during the project anyway.

GLMs
Mixed models
Flexible regression
Spatial statistics
Time series

Project 21: Evaluation of forensic toxicology data using classification methods

Statistics Supervisor: Diana Giurghita

External Supervisor (if any): NA

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Statistics is widely used in forensic toxicology applications requiring the analysis of data representing samples collected from crime scenes, such as: glass fragments, car paints, fire debris, voice recordings etc.. More specifically, classification methods are routinely used to identify the origin of collected samples by comparing a set of their descriptors with an existing database of known-origin samples.

This project will look at a data set of 125 individuals, classified as either chronic or non-chronic alcohol drinkers and 8 variables representing concentration values of direct and indirect biomarkers of ethanol consumption detected in blood or hair. According to the World Health Organisation, excessive alcohol consumption is a causal factor in more than 200 disease and injury conditions. Furthermore, the abuse of alcohol severely influences consumers' lives, leading to various legal, physical and psychological consequences, especially when dealing with behaviours that might cause road and work accidents.

The aim of the project is to apply different classification methods to determine the status of each individual in the data set based on the biomarkers collected. These methods are widely used in the forensic sciences, and some of them should already be familiar to the student, such as: linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbours (but there is a possibility to expand to other methods, such as: support-vector machines and decision trees if time permits). A comparison study will be performed between these methods and cross-validation will be employed to compute the performance of each method based on various indicators routinely used to determine the performance of classification algorithms (accuracy, sensitivity, specificity etc.). Finally, the project will report the best model based on the chosen indicators and will provide a more in-depth discussion about its expected performance when classifying future individuals as well as the most important predictors in the model.

Key Questions of Interest

Based on an initial exploratory analysis, are the variables collected informative about the status of each individual (chronic or non-chronic alcohol drinker)?

What classification methods are suitable for this data set?

How well do the models investigated fit? Are the model assumptions plausible?

What indicators can be used to compare different classification methods?

Which classification method performs the best? How good is the classification based on the available variables? What are the most important predictors in the best model?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The methods required in this project are already implemented in R, as such, the student is only required to read existing package documentation and apply relevant functions.

The student is expected to know how to perform basic data manipulations: dividing a data set into training, testing and validation sets and comparing and collating results from different runs.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Cross-validation, Binary classification methods (for example: LDA, QDA, knn, decision trees, support-vector machines etc.)

Project 22: A penalised logistic regression approach to analysing forensic toxicology data

Statistics Supervisor: Diana Giurghita

External Supervisor (if any): NA

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Statistics is widely used in forensic toxicology applications requiring the analysis of data representing samples collected from crime scenes, such as: glass fragments, car paints, fire debris, voice recordings etc.. More specifically, classification methods are routinely used to identify the origin of collected samples by comparing a set of their descriptors with an existing database of known-origin samples.

This project will look at a data set of 125 individuals, classified as either chronic or non-chronic alcohol drinkers and 8 variables representing concentration values of direct and indirect biomarkers of ethanol consumption detected in blood or hair. According to the World Health Organisation, excessive alcohol consumption is a causal factor in more than 200 disease and injury conditions. Furthermore, the abuse of alcohol severely influences consumers' lives, leading to various legal, physical and psychological consequences, especially when dealing with behaviours that might cause road and work accidents.

The aim of the project is to apply different penalised logistic regression methods to determine the status of each individual in the data set based on the biomarkers collected. These methods are widely used in the statistics and machine learning communities, but not so much in the context of forensic data analysis. While the most basic logistic regression model can be used as a classification method to predict the class of a binary response variable, this method fails when separation occurs in a dataset (i.e.: some predictors in the data set perfectly separate the classes in the response variable). In this situation, penalised logistic regression models, such as Firth GLM, Bayes GLM, or other regularized regression models (ridge, Lasso, elastic nets) provide more stable estimates, making them a more suitable choice. As such, another aim of this project is to compare the reliability and performance of the logistic regression model and a choice of a few penalised logistic regression models applied to the alcohol biomarkers dataset.

A comparison study will be performed between the chosen methods and cross-validation will be employed to compute the performance of each method based on various indicators routinely used to determine the performance of classification algorithms (accuracy, sensitivity, specificity etc.). Finally, the project will report the best model based on the chosen indicators and will provide a more in-depth discussion about its expected performance and reliability when classifying future individuals as well as the most important predictors in the model.

Key Questions of Interest

- Based on an initial exploratory analysis, are the variables collected informative about the status of each individual (chronic or non-chronic alcohol drinker)?
    - Does separation (perfect classification) or collinearity occur in this data set?

- What classification methods are suitable for this data set?
    - Fit a logistic regression model and comment on the model fit.
    - Fit a few penalised regression models (choosing from: Firth GLM, Bayes GLM, Lasso, Ridge or elastic net) and compare their fit with the logistic regression model.

- What indicators can be used to compare different classification methods?

- Perform a formal comparison study of the chosen penalised methods and indicate the best performing model.
    - How good is the classification based on the available variables?
    - What are the most important predictors?
    - Comment on the reliability of the chosen methods.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The methods required in this project are already implemented in R (glmnet, arm, brglm2), as such, the student is only required to read existing package documentation and apply relevant functions.

The student is also expected to know how to perform basic data manipulations such as dividing a data set into training, testing and validation sets and comparing and collating results from different runs.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Knowledge of generalized linear models (GLMs) (in particular logistic regression models) is essential. The student will learn about penalised logistic regression models, such as: Firth GLM, Bayes GLM, ridge, Lasso and elastic nets.

Project 23: A penalised approach to analysing forensic toxicology data using logistic regression fusion

Statistics Supervisor: Diana Giurghita

External Supervisor (if any): NA

Can be adapted to: Single / MSciWP

Brief Description of Project

Statistics is widely used in forensic toxicology applications requiring the analysis of data representing samples collected from crime scenes, such as: glass fragments, car paints, fire debris, voice recordings etc.. More specifically, classification methods are routinely used to identify the origin of collected samples by comparing a set of their descriptors with an existing database of known-origin samples. The main aim of the forensic analysts is to evaluate the physicochemical data from the collected evidence in the framework of two independent and mutually exclusive hypotheses. In recent years, the likelihood ratio (LR) has been widely adopted in forensic sciences and courtrooms since it expresses the strength of the observed evidence in favour of an existing hypothesis in a very straightforward way.

This project will look at a data set of 125 individuals, classified as either chronic or non-chronic alcohol drinkers and 8 variables representing concentration values of direct and indirect biomarkers of ethanol consumption detected in blood or hair. According to the World Health Organisation, excessive alcohol consumption is a causal factor in more than 200 disease and injury conditions. Furthermore, the abuse of alcohol severely influences consumers' lives, leading to various legal, physical and psychological consequences, especially when dealing with behaviours that might cause road and work accidents.

While the most basic logistic regression model can be used as a classification method to predict the class of a binary response variable, this method fails when separation occurs in a dataset (i.e.: some predictors in the data set perfectly separate the classes in the response variable). In this situation, penalised logistic regression models, such as Firth GLM, Bayes GLM, or other regularized regression models (ridge, Lasso, elastic nets) provide more stable estimates, making them a more suitable choice. Furthermore, logistic regression can also be used to produce a weighted average of LRs (called "scores") obtained using different pieces of evidence (or variables). These scores can be derived by first estimating the distribution of both classes for each of the variables of interest, and then calculating the LR (ratio of the probabilities) for a specific measurement under each class distribution. This approach, sometimes referred to as logistic regression fusion, is routinely used in forensic voice comparison and is especially useful for multivariate data when the number of variables is relatively large compared to the number of observations.

The main aim of the project is to demonstrate how different penalised logistic regression methods can be combined with the logistic regression fusion idea to construct a classification model. These models will then be applied to the dataset with the goal of predicting the status of each individual in the data set based on the LRs (or scores) of the biomarkers collected.

A secondary aim of this project is to compare the reliability and performance of the penalised logistic regression fusion models when applied to the alcohol biomarkers dataset. To this end, a comparison study will be performed between the chosen methods and cross-validation will be employed to compute the performance of each method based on various indicators routinely used to determine the performance of classification algorithms (accuracy, sensitivity, specificity etc.). Finally, the project will report the best model based on the chosen indicators and will provide a more in-depth discussion about its expected performance and reliability when classifying future individuals as well as the most important predictors in the model.

Key Questions of Interest

- Based on an initial exploratory analysis, are the variables collected informative about the status of each individual (chronic or non-chronic alcohol drinker)?
    - Does separation or collinearity (perfect classification) occur in this data set?

- What classification methods are suitable for this data set?
    - Calculate scores based on each predictor variable and fit a logistic regression fusion model. Comment on the model fit and identify and problem areas.
    - Combine the logistic regression fusion model and the penalised logistic regression idea to fit a penalised logistic regression fusion. The student can choose a few from the following penalized regression models: Firth GLM, Bayes GLM, Lasso, Ridge or elastic net. Compare their fit with the logistic regression fusion model.

- What indicators can be used to compare different classification methods?

- Perform a formal comparison study of the chosen penalised logistic regression fusion methods and indicate the best performing model.
    - How good is the classification based on the available variables?
    - What are the most important predictors?
    - Comment on the reliability of the chosen methods.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The methods required in this project are already implemented in R (glmnet, arm, brglm2), as such, the student is only required to read existing package documentation and apply relevant functions.

The student is expected to know how to perform basic data manipulations, such as: dividing a data set into training, testing and validation sets and comparing and collating results from different runs. More substantial programming will be required to write functions that convert the explanatory variables into scores.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Knowledge of generalized linear models (GLMs) (in particular logistic regression models) is essential. The student will learn about penalised logistic regression models, such as: Firth GLM, Bayes GLM, ridge, Lasso and elastic nets, and logistic regression fusion.

Project 24: Understanding the impact of global change – modelling growth of rainforest trees

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Prof. David Burslem

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Tropical forests are the most species-rich terrestrial plant communities on the planet, and retain globally significant reservoirs of carbon fixed in biomass. In many countries, tropical forests are being disturbed and degraded by timber exploitation, fire, over-hunting, fire and outright clearance for agricultural production. To respond to these pressures and develop a baseline for understanding how tropical forests respond to global change, forest ecologists have created a global network of large-scale forest dynamics plots in tropical forests, on which stems of all trees > 1 cm in diameter have been measured, identified and mapped. This network now comprises 71 sites in 27 countries, on which 6 million individual trees of approximately 12000 species are permanently tagged and monitored.

Through collaborators at the University of Aberdeen we have access to data from two censuses of a 50-ha forest dynamics plot at Danum Valley in Sabah, Malaysia, which includes about 250,000 trees of nearly 700 species. These data represent a rich resource for addressing questions of ecological importance, but the structure of the data-sets presents statistical challenges requiring careful handling and potentially new methods. Ecologists are keen to understand how the interactions among trees with their neighbours and the local environment allow them to persist through time, to test theories for the maintenance of species richness in plant communities. However, plant distributions are inherently spatially auto-correlated because of dispersal limitation, and therefore individual trees cannot be regarded as statistically independent. Recent statistical research has developed methods for accounting for spatially auto-correlated point pattern data, and these have been applied to investigating the relative importance of environmental covariates, such as nutrient availability and topography, on tree species distributions on the Danum plot. However, there has been no attempt to extend this research agenda to understanding how the dynamics of tree populations may be affected by environmental covariates within local neighbourhoods.

This project will focus on modelling **tree growth** within the Danum rainforest plot and is a collaborative project with Prof. David Burslem at the School of Biological Sciences, University of Aberdeen.

Key Questions of Interest

- Which local condition facilitate tree growth rates in tropical trees?
- Do growth rates in tropical trees vary among different tree species or families?
- Do these conclusions change when spatial structure/correlation is accounted for?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will use R to fit glm or gamm models as well as potentially various suitable spatial models.

The student will likely write some suitable functions to ease data handling and repeated model fitting.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models, Spatial statistics, Bayesian statistics

Project 25: Understanding the impact of global change – modelling mortality in rainforest trees

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Prof. David Burslem

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Tropical forests are the most species-rich terrestrial plant communities on the planet, and retain globally significant reservoirs of carbon fixed in biomass. In many countries, tropical forests are being disturbed and degraded by timber exploitation, fire, over-hunting, fire and outright clearance for agricultural production. To respond to these pressures and develop a baseline for understanding how tropical forests respond to global change, forest ecologists have created a global network of large-scale forest dynamics plots in tropical forests, on which stems of all trees > 1 cm in diameter have been measured, identified and mapped. This network now comprises 71 sites in 27 countries, on which 6 million individual trees of approximately 12000 species are permanently tagged and monitored.

Through collaborators at the University of Aberdeen we have access to data from two censuses of a 50-ha forest dynamics plot at Danum Valley in Sabah, Malaysia, which includes about 250,000 trees of nearly 700 species. These data represent a rich resource for addressing questions of ecological importance, but the structure of the data-sets presents statistical challenges requiring careful handling and potentially new methods. Ecologists are keen to understand how the interactions among trees with their neighbours and the local environment allow them to persist through time, to test theories for the maintenance of species richness in plant communities. However, plant distributions are inherently spatially auto-correlated because of dispersal limitation, and therefore individual trees cannot be regarded as statistically independent. Recent statistical research has developed methods for accounting for spatially auto-correlated point pattern data, and these have been applied to investigating the relative importance of environmental covariates, such as nutrient availability and topography, on tree species distributions on the Danum plot. However, there has been no attempt to extend this research agenda to understanding how the dynamics of tree populations may be affected by environmental covariates within local neighbourhoods.

This project will focus on modelling **mortality** within the Danum rainforest plot and is a collaborative project with Prof. David Burslem at the School of Biological Sciences, University of Aberdeen.

Key Questions of Interest

- Which local condition facilitate mortality in tropical trees?
- Does mortality in tropical trees vary among different tree species or families?
- Do these conclusions change when spatial structure/correlation is accounted for?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will use R to fit glm or gamm models as well as potentially various suitable spatial models.

The student will likely write some suitable functions to ease data handling and repeated model fitting.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models, Spatial statistics, Bayesian statistics

Project 26: Understanding the impact of global change – modelling recruitment of new trees in rainforest trees

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Prof. David Burslem

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Tropical forests are the most species-rich terrestrial plant communities on the planet, and retain globally significant reservoirs of carbon fixed in biomass. In many countries, tropical forests are being disturbed and degraded by timber exploitation, fire, over-hunting, fire and outright clearance for agricultural production. To respond to these pressures and develop a baseline for understanding how tropical forests respond to global change, forest ecologists have created a global network of large-scale forest dynamics plots in tropical forests, on which stems of all trees > 1 cm in diameter have been measured, identified and mapped. This network now comprises 71 sites in 27 countries, on which 6 million individual trees of approximately 12000 species are permanently tagged and monitored.

Through collaborators at the University of Aberdeen we have access to data from two censuses of a 50-ha forest dynamics plot at Danum Valley in Sabah, Malaysia, which includes about 250,000 trees of nearly 700 species. These data represent a rich resource for addressing questions of ecological importance, but the structure of the data-sets presents statistical challenges requiring careful handling and potentially new methods. Ecologists are keen to understand how the interactions among trees with their neighbours and the local environment allow them to persist through time, to test theories for the maintenance of species richness in plant communities. However, plant distributions are inherently spatially auto-correlated because of dispersal limitation, and therefore individual trees cannot be regarded as statistically independent. Recent statistical research has developed methods for accounting for spatially auto-correlated point pattern data, and these have been applied to investigating the relative importance of environmental covariates, such as nutrient availability and topography, on tree species distributions on the Danum plot. However, there has been no attempt to extend this research agenda to understanding how the dynamics of tree populations may be affected by environmental covariates within local neighbourhoods.

This project will focus on modelling **recruitment of new trees** (appearance of a new individual above the minimum size at the second census) within the Danum rainforest plot and is a collaborative project with Prof. David Burslem at the School of Biological Sciences, University of Aberdeen.

Key Questions of Interest

- Which local condition facilitate **recruitment** in tropical trees?
- Does **recruitment** in tropical trees vary among different tree species or families?
- Do these conclusions change when spatial structure/correlation is accounted for?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will use R to fit glm or gamm models as well as potentially various suitable spatial models.

The student will likely write some suitable functions to ease data handling and repeated model fitting.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models, Spatial statistics, Bayesian statistics

Project 27: Assessing gender imbalance in citation practices in scientific publications in mathematics

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Many studies have found large and persistent imbalances in academic participation among the different genders across all scientific fields. These may be quantified and hence highlighted in terms of obvious measures such as the proportion of active female scientists in a field at different levels in the academic hierarchy. However, other measures of academic inclusion and success may also be considered to gain an improved understanding of the current state of these imbalances – and to alleviate these issues. One particular measure that is often used to assess an academic's reputation and the relevance of their work is the number of times their publications are cited by their colleagues. A gender imbalance in citation levels would hence potentially have a negative effect on the career perspectives of women.

In a recent paper in the well-respected scientific journal "Nature Neurosciences", Dworking et al. (2020) conduct a detailed statistical analysis of citation practices in the Neurosciences. They analyse citation data from five top neuroscience journals and find that reference lists tend to include more papers with men as senior authors. In addition, they show that this overcitation of men and undercitation of women is driven largely by the citation practices of men, and that it is increasing over time as the field becomes more diverse.

This project will focus on assessing potential imbalances in citation practice within mathematics and will adapt the existing study in Dworking et al. (2020) to citations in the field of mathematics.

Dworkin, J.D., Linn, K.A., Teich, E.G. *et al.* The extent and drivers of gender imbalance in neuroscience reference lists. *Nat Neurosci* **23,** 918–926 (2020). https://doi.org/10.1038/s41593-020-0658-y

Key Questions of Interest

- How can we best adapt the existing study to the field of mathematics?
- Do papers in top mathematics journal cite papers by men mode frequently than by women?
- Do we find a similar pattern in mathematics as has been identified by Dworking et al (2020) of the over-/undercitation being driven by men?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will use existing R code provided by Dworking et al. (2020) but might have to make changes to the code when adapting the approach for their use.

The student will likely write some suitable functions to ease data handling etc.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models

Project 28: Assessing gender imbalance in citation practices in scientific publications in statistics

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Many studies have found large and persistent imbalances in academic participation among the different genders across all scientific fields. These may be quantified and hence highlighted in terms of obvious measures such as the proportion of active female scientists in a field at different levels in the academic hierarchy. However, other measures of academic inclusion and success may also be considered to gain an improved understanding of the current state of these imbalances – and to alleviate these issues. One particular measure that is often used to assess an academic's reputation and the relevance of their work is the number of times their publications are cited by their colleagues. A gender imbalance in citation levels would hence potentially have a negative effect on the career perspectives of women.

In a recent paper in the well-respected scientific journal "Nature Neurosciences", Dworking et al. (2020) conduct a detailed statistical analysis of citation practices in the Neurosciences. They analyse citation data from five top neuroscience journals and find that reference lists tend to include more papers with men as senior authors. In addition, they show that this overcitation of men and undercitation of women is driven largely by the citation practices of men, and that it is increasing over time as the field becomes more diverse.

This project will focus on assessing potential imbalances in citation practice within statistics and will adapt the existing study in Dworking et al. (2020) to citations in the field of statistics.

Dworkin, J.D., Linn, K.A., Teich, E.G. *et al.* The extent and drivers of gender imbalance in neuroscience reference lists. *Nat Neurosci* **23,** 918–926 (2020). https://doi.org/10.1038/s41593-020-0658-y

Key Questions of Interest

- How can we best adapt the existing study to the field of statistics?
- Do papers in top statistics journal cite papers by men mode frequently than by women?
- Do we find a similar pattern as has been identified by Dworking et al (2020) of the over-/undercitation being driven by men?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will use existing R code provided by Dworking et al. (2020) but might have to make changes to the code when adapting the approach for their use.

The student will likely write some suitable functions to ease data handling etc.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models

Project 29: Analysis of seabird behaviour based on GPS tracks

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Dr. Sarah Saldanha/Fernando Medrano (Barcelona)

Can be adapted to: Single/ MSciWP (delete only if necessary)

Seabirds are marine top predators breeding on land but feeding at sea. During breeding, they perform foraging trips from their nests on land to prey-rich areas at sea to feed, and back to the nest to incubate the egg or feed the chick. Their foraging behaviour is especially sensitive to environmental changes, so their foraging movements can be used as indicators of oceanic environment health. By fitting seabirds with electronic devices that record their position we can understand their relationship to the environment through their distributions and behaviour.

The analysis of bird movement data aims to distinguish different types of behaviour of the bird, based on turning angles and speed, where high speed and small angles indicate directional flight, whereas slow speed and large turning angles indicate prey searching behaviour.

In this project we will analyse GPS tracking data of individual seabirds. These trackers collect data at low temporal resolution (1 hour or more between fixes) and often at irregular points in time since the device cannot always record data. Due to the low temporal resolution, bird behaviour cannot be derived from turning angles and speed, since consecutive locations are too far apart to tell us much about the bird's behaviour.

The project will use hidden Markov models to attempt classification to derive bird behaviour from the tracking data. This is a joint project with the Seabird Ecology research group of the University of Barcelona and the IRBio (Institute for the Research of Biodiversity).

Key Questions of Interest

- Can Hidden Markov Models be used to analyse GPS tracking data of seabirds?
- Can we distinguish different types of behaviours using Hidden Markov models on these data or is the temporal resolution too coarse for this purpose?
- What ecological interpretations can we get from the analysis of these data?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will use and explore software packages developed for the analysis of bird tracking data using Hidden Markov models.

The student will likely write some suitable functions to ease data handling etc.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models

Project 30: Visualising high-dimensional data of seabird behaviour

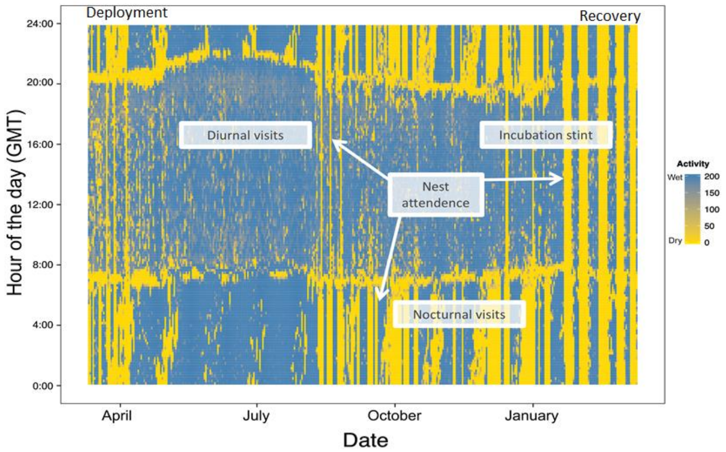
Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Dr. Virginia Morera Pujol (Barcelona)

Can be adapted to: Single/ MSciWP (delete only if necessary)

Seabirds are marine top predators breeding on land but feeding at sea. During breeding, they perform foraging trips from their nests on land to prey-rich areas at sea to feed, and back to the nest to incubate the egg or feed the chick, and during the non-breeding period they stay constantly at sea. Since they usually spend the non-breeding season in remote areas far away from the colonies, one of the only ways to analyse their behaviour is to deploy sensors that detect the water connectivity, allowing us to know whether the birds is on the water (wet) or on air/land (dry).

This project focuses on data visualisation based on heatmaps similar to the figure shown below, in order to represent behaviour of a bird along the year. Here days are displayed on the horizontal axis and hours of the day on the vertical one. Blue values indicate large proportions of time were spent on the water per hour, yellow values mean large proportions of time were spent flying or on land, i.e. near at the burrow.



This is used to identify specific behaviours such as days where the bird stays at the nest (dry, and thus, yellow) all day, so they in the breeding colonies. Wider yellow bars mean more days spent at the nest incubating the egg, when the adult does not leave the nest until the partner comes back, while wide blue sections identify periods when the bird spent long hours on the water (for example, when birds are moulting their feathers). In this way ecologist can detect changes in activity of the bird. However, these visualizations only represent one bird at a time. This project looks into way of representing this information this for an entire population.

This is a joint project with the Seabird Ecology research group of the University of Barcelona and the IRBio (Institute for the Research of Biodiversity).

Key Questions of Interest

- What properties does a good visualisation technique have?
- What are specific challenges in the context of visualising seabird tracking data through time?
- What visualisation methods can be used to visualise the behaviour of more than one bird/ and entire colony?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will explore different modern visualisation methods and will have to use and explore the associated R packages for this purpose.

The student will likely write some suitable functions to ease data handling etc.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models

Project 31: Analysis of seabird behaviour based on multivariate accelerometer data

Statistics Supervisor: Prof. Janine Illian

External Supervisor (if any): Leia Navarro Herrero (Barcelona)

Can be adapted to: Single/ MSciWP (delete only if necessary)

Seabirds are marine top predators breeding on land but feeding at sea. During breeding, they perform foraging trips from their nests on land to prey-rich areas at sea to feed, and back to the nest to incubate the egg or feed the chick. Their foraging behaviour is especially sensitive to environmental changes, so their foraging movements can be used as indicators of oceanic environment health. By fitting seabirds with electronic devices that record their position, and with additional sensors measuring, for example, tri-axial acceleration, we can understand their relationship to the environment through their distributions and behaviour.

The analysis of bird movement data aims to distinguish different types of behaviour of the bird, based on turning angles and speed, where high speed and small angles indicate directional flight, whereas slow speed and large turning angles indicate prey searching behaviour.

This project however, analyses **accelerometer data** representing acceleration along three axes (x, y and z) at 25Hz. These data, together with the position data can be used to identify behaviour, such as diving, energy budgets or habitat use.
This project will use multivariate classification methods that have been specifically developed or the use with this kind of data. In particular, they will use a package called bigMap that uses parallelized t-SNE to classify, in this case, behaviour, according to accelerometer and location data as discussed in Garriga, J., & Bartumeus, F. (2018). bigMap: big data mapping with parallelized t-SNE. *arXiv preprint arXiv:1812.09869*.

This is a joint project with the Seabird Ecology research group of the University of Barcelona and the IRBio (Institute for the Research of Biodiversity).

Key Questions of Interest

- What are specific challenges in the context of accelerometer and location data?
- How does the approach in bigMap classify accelerometer and location data?
- What results do we get when the approach is applied to real data such as xxx?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will explore classification methods and will have to use and explore the associated R packages for this purpose.

The student will likely write some suitable functions to ease data handling etc.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models , Multivariate methods
Desirable – Linear Mixed Models

Project 32: Spatio-temporal modelling of coronary heart disease in Scotland

Statistics Supervisor: Eilidh Jack

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

The field of disease mapping focuses on showing the extent to which the risk of disease varies across space. Different regions often have different risk levels, and this is generally a result of differences in environmental and socio-economic factors between regions. This is particularly true in Scotland, which has the widest health inequalities in Western Europe. On average, men living in the most affluent areas experience 23.8 more years of good health than those living in the most deprived (22.6 for women).

Much of the modern methodology for estimating disease is based on conditional autoregressive (CAR) models. These models allow for spatial correlation between neighbouring areas, based on the idea that nearby areas are likely to have more in common than those which are further apart. These CAR models can also be extended into spatio-temporal models which account for changes in the disease risk surface over time.

In this project, you will have the opportunity to explore the spatio-temporal pattern of risk across Scotland for coronary heart disease between 2002 and 2012. You will need to identify and fit an appropriate spatio-temporal model to these data, and then produce maps of the risk across the study period. This will allow you to investigate the trends in disease risk over time, and to identify the highest and lowest risk areas.

Key Questions of Interest

- How does the risk of coronary heart disease vary across Scotland?
- Does this spatial pattern of risk change over time?
- Which regions have the highest and lowest risk of disease? Does this change over time?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will learn to fit spatial models and plot disease maps using R.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Linear Mixed Models, Spatial statistics, Time series.

Project 33: Estimating Cholera risk in endemic countries

Statistics Supervisor: Eilidh Jack

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Cholera is an acute diarrhoeal disease that can be deadly if left untreated. Despite being an easily treatable disease Cholera remains a global threat to public health and is an indicator of inequity and lack of social development. Cholera can be endemic or epidemic, with endemic areas being identified as areas where cholera cases were detected in the last three years and the disease is spread through local transmission.

In this project, you will have the opportunity to explore the spatial pattern of cholera risk in countries across the world where cholera is an endemic disease. You will need to identify and fit an appropriate spatial model to these data, and then produce maps of the risk. This will allow you to investigate spatial patterns in cholera risk and to identify the highest and lowest risk countries. You will also have access to several covariates, including information on access to clean water and sanitation, which can be used to help explain risk of cholera.

Key Questions of Interest

- How does the risk of cholera vary in endemic countries across the world?
- What effect do the covariates have on risk of cholera?
- Which endemic counties have the highest and lowest risk of cholera?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will learn to fit spatial models and plot disease maps using R.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised Linear Models
Desirable – Linear Mixed Models, Spatial Statistics

Project 34: Estimating changes in health inequalities across three of Scotland's biggest killers

Statistics Supervisor: Eilidh Jack

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Disease risk can often vary substantially across a region or a country as a result of socio-economic inequalities. This is particularly true in Scotland, which has the widest health inequalities in Western Europe. On average, men living in the most affluent areas experience 23.8 more years of good health than those living in the most deprived (22.6 for women).

In this project, you will have the opportunity to explore the spatial patterns of risk across Scotland for three major diseases: Coronary Heart Disease, Cerebrovascular Disease and Respiratory Disease. You will need to identify and fit an appropriate model to these data, and then produce maps of the risk for each disease across the study region. This will allow you to investigate the spatial patterns in risk which occur across Scotland for different diseases, and to identify the highest and lowest risk areas.

Key Questions of Interest

- How does the pattern of disease risk in Scotland vary between different diseases?
- Which areas have the highest and lowest risk for each disease and are these common across all diseases?
- How does inequality in disease risk differ between different diseases?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will learn to fit spatial models and plot disease maps using R.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised Linear Models
Desirable – Linear Mixed Models, Spatial Statistics, Multivariate Methods

Project 35: Estimating changes in health inequalities in respiratory disease across Scotland

Statistics Supervisor: Eilidh Jack

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Disease risk is not constant over space and time and is often impacted by exposure to risk inducing behaviour such as consumption of alcohol. Poverty, and more generally deprivation, are major factors in the spatial variation observed in the risk of disease, with more highly deprived areas usually exhibiting elevated levels of disease risk. This difference in disease risk between social groups and population areas is known as a health inequality. Scotland has the widest health inequalities in Western Europe. On average, men living in the most affluent areas experience 23.8 more years of good health than those living in the most deprived (22.6 for women).

In this project, you will have the opportunity to explore the spatio-temporal pattern of risk across Scotland for respiratory disease between 2002 and 2012, and thus estimate how health inequalities are changing over time for this disease. You will need to identify and fit an appropriate spatio-temporal model to these data, and then produce maps of the risk across the study period. This will allow you to investigate the trends in disease risk over time, and to identify the highest and lowest risk areas.

Key Questions of Interest

- How does the risk of respiratory disease vary across Scotland?
- How are health inequalities changing over time in Scotland for respiratory disease risk?
- Which regions have the highest and lowest risk of disease?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will learn to fit spatial models and plot disease maps using R.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised Linear Models
Desirable – Linear Mixed Models, Spatial Statistics, Time series.

Project 36: Spatio-temporal analysis of respiratory disease in primary care.

Statistics Supervisor: Duncan Lee

External Supervisor (if any): NA

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Disease mapping is the area of spatial epidemiology concerned with visualising and quantifying the spatio-temporal variation in disease risk. Key questions in such analyses typically include: (i) where are the areas at highest risk of disease; and (ii) what factors affect disease risk? Answering both these questions allow epidemiologists and public health professionals to target health-based interventions where they are most needed, such as an advertising campaign about potential risk factors for the disease under study. Most of these studies model data relating to severe cases of ill health, such as hospitalisation or death. However, there is a large amount of ill health amongst the population that is less severe, and is instead treated in primary (non-hospitalised) care by visiting your general practice (GP) doctor for medication (called a prescription).

This project use population level data relating to 724 GP surgeries in Scotland on a monthly basis between January 2016 and December 2019. We focus on respiratory disease in this project, such as asthma and chronic obstructive pulmonary disease (COPD). Data on the prevalence of these conditions is not available, so instead we use data on the numbers of prescriptions for medicines used to treat these conditions as a proxy measure of the prevalence of disease. Specifically, the response variable is a count of the number of prescriptions written by each GP surgery in each month for medicines used to relieve the symptoms of asthma and chronic obstructive pulmonary disease (COPD) called short acting beta-2 agonists (e.g. Ventolin). Data on the expected numbers of prescriptions are also available based on the size and age/sex demographics of the GP surgery's patient population, which is computed using indirect standardisation. Finally, data on a number of covariate risk factors are available, including measures of meteorology, air pollution and socio-economic deprivation (poverty).

Key Questions of Interest

There are a number of possible questions of interest that one could answer for these data, including:

1. Which covariate factors affect the prevalence of respiratory ill health in primary care and what are their effects?

2. Which GP surgeries have high rates of respiratory prescribing, and has this set of surgeries remained consistent over the 4 years of the study?
3. What are the temporal trends in respiratory prescribing rates between 2016 and 2019, and has the level of spatial variation in these rates, often termed a health inequality, changed over time?
4. Do these data contain residual spatial and or temporal correlation, and if so how should this residual correlation be modelled?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***No***

If 'Yes', please specify what this might involve:

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential – Generalised linear models
Desirable – Biostatistics, Spatial statistics, Time series.

Project 37: Investigating parameter inference performance on systems described by differential equations using spline interpolation (A)

Statistics Supervisor: Benn Macdonald

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

Inferring the parameters of these equations can be achieved in a two-step approach. First, smooth the data with some interpolation method (splines interpolation in this case) in order to avoid observational error distorting the estimation. From this smooth interpolant, gradients at given points can be obtained. Second, the ODEs themselves will return gradients for a given parameter set and a penalised likelihood can be obtained by comparing the predicted gradients from the interpolant to those predicted by the ODEs. This is known as gradient matching. Statistical inference can be carried out by non-linear optimisation or MCMC.

Splines based interpolation depends on several tuning parameters e.g. the number of knots to set, the polynomial order for the spline and the penalty parameter for the roughness penalty (if penalised splines are used, for example). In the two-step approach described above, the parameter estimation will depend on how well the interpolant is able to model the underlying true function that the data comes from. The goal of this project is to investigate how the inference performance is affected by the various factors that the interpolant depends on.

Key Questions of Interest

- How does the accuracy of parameter estimation depend on the number of knots?
- How does the accuracy of parameter estimation depend on the polynomial order?
- How does the accuracy of parameter estimation depend on the roughness penalty?
- Is this consistent across dataset size?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then performing optimisation/sampling depending on the student's preference for Frequentist/Bayesian Statistics.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Non-linear optimisation and spline interpolation are required for the project.

Project 38: Investigating parameter inference performance on systems described by differential equations using spline interpolation (B)

Statistics Supervisor:

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

Inferring the parameters of these equations can be achieved in a two-step approach. First, smooth the data with some interpolation method (splines interpolation in this case) in order to avoid observational error distorting the estimation. From this smooth interpolant, gradients at given points can be obtained. Second, the ODEs themselves will return gradients for a given parameter set and a penalised likelihood can be obtained by comparing the predicted gradients from the interpolant to those predicted by the ODEs. This is known as gradient matching. Statistical inference can be carried out by non-linear optimisation or MCMC.

The accuracy of the estimation will depend on the noise of the observed system. As the signal-to-noise (SNR) level gets smaller for a system, the underlying function becomes more difficult to model (since there is more error in the data). The goal of this project is to investigate how the inference performance is affected by the signal-to-noise ratio and whether this is consistent across ODE system and complexity.

Key Questions of Interest

- As the SNR decreases, can we quantify the decrease in parameter inference accuracy?
- Is this similar for different observed signals in a given ODE model i.e. for different “true” parameter sets for the same ODE model?
- Is this consistent across different ODE model?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then performing optimisation/sampling depending on the student’s preference for Frequentist/Bayesian Statistics.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Non-linear optimisation and spline interpolation are required for the project.

Project 39: Investigating parameter inference performance on systems described by differential equations using Gaussian process interpolation (A)

Statistics Supervisor: Benn Macdonald

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

Inferring the parameters of these equations can be achieved in a two-step approach. First, smooth the data with some interpolation method (Gaussian process (GP) interpolation in this case) in order to avoid observational error distorting the estimation. From this smooth interpolant, gradients at given points can be obtained (in fact, for a GP a distribution over the gradients can be obtained in closed form). Second, the ODEs themselves will return gradients for a given parameter set and a penalised likelihood can be obtained by comparing the predicted gradients from the interpolant to those predicted by the ODEs. This is known as gradient matching. Statistical inference can be carried out by non-linear optimisation or MCMC.

Gaussian processes are very flexible and able to model highly non-linear functions. They depend on a kernel (which can be chosen to reflect prior knowledge of the modelled system), which depends on hyperparameters. In the two-step approach described above, the parameter estimation will depend on how well the interpolant is able to model the underlying true function that the data comes from. The goal of this project is to investigate how the inference performance is affected by the hyperparameters for a given kernel.

Key Questions of Interest

- How does the accuracy of parameter estimation depend on the choice of hyperparameters?
- Is this consistent across dataset size?
- Is the trend dependent on the level of observational noise?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then performing optimisation/sampling depending on the student’s preference for Frequentist/Bayesian Statistics.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Non-linear interpolation via Gaussian processes and non-linear optimisation/MCMC is required for this project.

A student is unlikely to have covered Gaussian processes and so will spend time learning this approach.

Project 40: Investigating parameter inference performance on systems described by differential equations using Gaussian process interpolation (B)

Statistics Supervisor: Benn Macdonald

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

Inferring the parameters of these equations can be achieved in a two-step approach. First, smooth the data with some interpolation method (Gaussian process (GP) interpolation in this case) in order to avoid observational error distorting the estimation. From this smooth interpolant, gradients at given points can be obtained (in fact, for a GP a distribution over the gradients can be obtained in closed form). Second, the ODEs themselves will return gradients for a given parameter set and a penalised likelihood can be obtained by comparing the predicted gradients from the interpolant to those predicted by the ODEs. This is known as gradient matching. Statistical inference can be carried out by non-linear optimisation or MCMC.

Gaussian processes are very flexible and able to model highly non-linear functions. They depend on a kernel (which can be chosen to reflect prior knowledge of the modelled system), which depends on hyperparameters that can be inferred from the data. In the two-step approach described above, the parameter estimation will depend on how well the interpolant is able to model the underlying true function that the data comes from. The goal of this project is to investigate how the inference performance is affected by the choice of kernel.

Key Questions of Interest

- How does the accuracy of parameter estimation depend on the choice of kernel?
- Is this similar for different observed signals in a given ODE model i.e. for different “true” parameter sets for the same ODE model?
- Is this consistent across different ODE model?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then performing optimisation/sampling depending on the student’s preference for Frequentist/Bayesian Statistics.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Non-linear interpolation via Gaussian processes and non-linear optimisation/MCMC is required for this project.

A student is unlikely to have covered Gaussian processes and so will spend time learning this approach.

Project 41: Parameter inference for systems described by differential equations: Explicit solution vs Gradient matching

Statistics Supervisor: Benn Macdonald

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

The parameters of these equations can be inferred using different approaches and this project will look at how inference using an explicit solution of the ODEs compares to that of gradient matching. The explicit solution approach involves solving the ODEs for a given set of initial conditions and parameters. This solution is compared to the data in order to calculate the likelihood. On the other hand, gradient matching first smooths the data with some interpolation method and from this smooth interpolant, gradients at given points can be obtained. The ODEs themselves will return gradients for a given parameter set and a penalised likelihood can be obtained by comparing the predicted gradients from the interpolant to those predicted by the ODEs. Inference can then be carried out by maximising the likelihood/penalised likelihood or sampling from the posterior.

The goal of this project will be to investigate how the inference performance compares between these two approaches.

Key Questions of Interest

- What is the accuracy of prediction using the explicit solution approach?
- What is accuracy of prediction using gradient matching?
- Is this consistent across observational noise level?
- Is this consistent across ODE model?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then performing optimisation/sampling depending on the student's preference for Frequentist/Bayesian Statistics.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Non-linear interpolation and non-linear optimisation will be required for this project.

Project 42: Model selection for systems described by differential equations

Statistics Supervisor: Benn Macdonald

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Ordinary differential equations (ODEs) are a powerful way of modelling observed systems, effectively providing a process with a mathematical description. Examples in biology include predator-prey dynamics in ecology (Lotka (1932)), autocatalysis in chemical kinetics (Atkins (1938)), cardiac excitation (Biktashev et al. (2008), Adon et al. (2015)) and kinetics of enzyme reactions (Gratie et al. (2013)).

Parameter inference can be carried out by solving the system of equations for a given parameter set and minimising the discrepancy between the predicted signals from the ODEs and the data. Since solutions to the ODEs typically do not exist in closed form, explicit solutions of the ODEs need to be computed numerically.

Model selection for ODEs aims at distinguishing between different hypotheses describing the structure of the systems. There are two main approaches to model selection - explanatory model selection and predictive model selection. Explanatory model selection is the method of integrating over the parameters and focussing on the marginal likelihood of the data i.e. the probability of the data given the model and not the probability of the data given some parameter set. The latter approach, predictive model selection, is a measure of out of sample predictive performance.

The goal of this project will be to compare different model selection methods for systems described by ODEs.

Key Questions of Interest

- How can we calculate the probability of the data given the model for systems described by differential equations?

- How do approaches such as AIC, BIC, cross validation, to name a few, compare in terms of model selection performance for systems described by ODEs?

- Is this consistent across dataset size?

- Is this dependent on the level of observational error?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The student will be required to use R, simulating data from a set of differential equations and then implement code that conducts Bayesian inference.

Typical plotting and data exploration skills in R will also be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

This project will involve a technique known as thermodynamic integration. The student is unlikely to have encountered this before and will spend time learning the method. Code will be provided in order to conduct this approach.

Knowledge of model selection approaches such as AIC, BIC, cross validation will be required for this project.

Project 43: Investigating how missing values affect predictive performance

Statistics Supervisor: Benn Macdonald

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Missing values in a dataset can result in a loss of information, either reducing a method’s power and ability to generalise to the wider population or introducing bias into subsequent conclusions. Missing values can occur completely at random (MCAR), at random (MAR) or not at random (MNAR), with the latter usually being the most problematic in practice.

In the presence of missing data (yes, the oxymoron was intentional) we can impute the missing values based on the remaining observed data. When the data are missing at random, this can often lead to unbiased results. However, this is not usually the case when the data are missing not at random.

The goal of this project will be to investigate the effect of the proportion of missing values in the dataset on the performance of out-of-sample prediction.

Key Questions of Interest

- How does the proportion of missing values affect the out-of-sample prediction accuracy?
- How can we impute the missing values and does this improve the out of sample prediction accuracy?
- Does this differ when the data are missing not at random?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Typical plotting, data exploration skills and programming in R will be essential.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Linear modelling and implementation of imputation/multiple imputation. The student is unlikely to have encountered the latter before and will spend time learning the approach.

Time permitting, Gaussian process interpolation could be employed as an alternative to multiple imputation. Again, students are unlikely to have encountered this before and would therefore need to spend time familiarising themselves with the method.

Project 44: Modelling Energy Intake with the National Diet and Nutrition Survey (NDNS)

Statistics Supervisor: Claus Mayer & Vlad Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Excessive Energy Intake (EI) is an obvious driver of obesity, one of the main health concerns in the world today, contributing to type 2 diabetes and cardiovascular disease among other things. In this project, we want to investigate which factors are associated with EI and thus may help us to predict it in the UK population. We will use data from National Diet and Nutrition Survey (NDNS), a rolling programme that collects data from around 1000 people per year on their food consumption and nutrient intake. Within the NDNS each individual's food intake is recorded on four consecutive days and there is also detailed information available about participants (age, gender, BMI, physical activity, household income, region etc…). The aim of the project is to model personal EI by these subject specific measurements.

1. We would expect a mean-variance relationship in energy intake measurements. Can we overcome this by a suitable transformation (eg. log) or could we use alternative methods to linear modelling like quantile regression to address this issue?
2. What are the key variables associated with EI and potential drivers of it?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

The data from different phases of the rolling program are stored in separate files. The format of these files will have changed in places so merging requires some care. Data on individuals are stored in separate files too and need to be integrated. All of this will need not so much programming but at least some data handling within R, before we conduct the analysis itself.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

The main analysis tool will be a standard linear model, possibly a mixed model if we want to model energy intake on a daily basis rather than the 4-day-average. Exploratory analysis might involve some multivariate analysis (principal components, hierarchical clustering). Depending on how fast the project progresses we might also investigate quantile regression as an alternative to transforming the response variable.

Project 45: How many authors do you need to write a statistics paper?

Statistics Supervisor: Claus Mayer & Alexey Lindo

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

The field of statistics has changed dramatically over the last few decades. For large parts of the 20th century limited computational resources required statistical methods to be expressed in closed formulas (if possible), thus making the development of such methods very much a mathematical task. With the increase of computing power alternative approaches like resampling or simulation based methods became feasible and gained popularity. At the same time, technological developments in other sciences gave rise to data that are far more complex in both structure and volume than previously known. In the modern age of Big Data developing new statistical methods thus needs a far more diverse skill set often requiring the collaboration of statisticians, computer scientists, numerical mathematicians and experts from the corresponding application area (eg. molecular biology, ecology, medicine etc.),

In this project, we want to study what effect this need for more collaboration has had on the number of authors of a statistics article. Importing Web of Science literature searches into R using the *bibliometrix* package allows us to create data sets that show the number of authors for each article in a given statistics journal over, say, the last 60 years. Based on these data we want to assess whether there has been any trend over time, when changes have occurred and whether the observed patterns are consistent across different journals.

1. What pattern does the number of authors on statistics paper show over time and how does this vary across different journals?
2. Can the number of authors be modelled by a Poisson distribution or do we need alternatives that can handle over or under-dispersion (negative binomial, generalized poisson)?
3. Can we model a smooth trend function for average number of authors over time with a general additive model (GAM)?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

The generation of the data set will need some programming as various reports for searches on the Web of Science will need to be brought together and the data of interest need to be extracted from these files. If the data turn out not to be compatible with a standard statistical model some additional programming might be necessary to use alternatives (e.g. resampling methods).

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

The main analysis tool will be a generalized linear model. A first easy step might be to dichotomize the outcome (single-author vs multi-author papers) and use a binomial model but ultimately we want to use the full count data. This is likely to need some kind of extension of a Poisson model (generalized Poisson) to address potential under-dispersion. General additive models (GAM) will be needed to model smooth trends across years.

Project 46: Climate change impacts on lake water quality

Statistics Supervisor: Claire Miller

External Supervisor: Yvonne McElarney, AFBI, Northern Ireland

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Eutrophication is a major threat to freshwater systems and over-enrichment of lakes can lead to undesirable effects on water quality. Water quality across Europe continues to decline (European Environment Agency, 2018) despite management measures implemented under the European Water Framework Directive (WFD) (European Commission, 2000). This project will explore the water quality for Lough Neagh, a large lake in Northern Ireland covering 383km^2, which is important for freshwater supplies to Northern Ireland and commercial fishing. The lake has been sampled manually twice a month by the Agri-food and Biosciences Institute (AFBI), in Northern Ireland and hence time series data are available for water chemistry, temperature, pH and depth, fortnightly since the mid-1970s. The focus of interest for this project is the long-term changes and seasonal patterns in water chemistry over the past 40 years, and specifically the impacts of temperature on changes in water chemistry. The identification of these patterns and trends in nutrient dynamics and lake ecology is necessary in order to effectively manage this lake.

Please Note: The data are confidential and hence will be covered by a data licence. This will state that the data are only to be used by the student for this project and cannot be transferred to another party.

Key Questions of Interest

For the different water chemistry variables:

1. What are the long-term trends?
2. Is there evidence of seasonal patterns?
3. Is there evidence of temperature impacts on changes in the water chemistry?
4. How do patterns and trends differ by depth?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount of programming may be required for data manipulation

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Data transformations
Generalised linear models
Flexible regression - (Generalised) additive models
Time series – autocorrelation/partial autocorrelation functions

Missing data imputation, techniques to deal with non-detects (survival analysis) and structural equation models are also possibilities for investigation, time permitting.

Project 47: Invasion of the zebra mussels

Statistics Supervisor: Claire Miller

External Supervisor: Yvonne McElarney, AFBI, Northern Ireland

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Eutrophication is a major threat to freshwater systems and over-enrichment of lakes can lead to undesirable effects on water quality. Water quality across Europe continues to decline (European Environment Agency, 2018) despite management measures implemented under the European Water Framework Directive (WFD) (European Commission, 2000). This project will explore the water quality for Lough Erne, a lake in Northern Ireland situated in County Fermanagh with a surface area of approximately 110 km^2; the lake is popular for angling and water sports. The lake has been sampled at 5 sites manually twice a month by the Agri-food and Biosciences Institute (AFBI), in Northern Ireland and hence time series data are available for water chemistry and the invasive species of zebra mussels, fortnightly since the early 1980s. The focus of interest for this project is the long-term changes and seasonal patterns in water chemistry, and the impact of zebra mussels. The identification of these trends in nutrient dynamics and lake ecology is necessary in order to effectively manage this lake.

Please Note: The data are confidential and hence will be covered by a data licence. This will state that the data are only to be used by the student for this project and cannot be transferred to another party.

Key Questions of Interest

For the different water chemistry variables:

1. What are the long-term trends?
2. Is there evidence of seasonal patterns?
3. Is there evidence of zebra mussel impacts on changes in the water chemistry?
4. How do patterns and trends differ by site?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount of programming may be required for data manipulation

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Data transformations
Generalised linear models
Flexible regression - (Generalised) additive models
Time series – autocorrelation/partial autocorrelation functions

Missing data imputation and techniques to deal with non-detects (survival analysis) are also possibilities for investigation, time permitting.

Project 48: Catchment impacts on lake water quality

Statistics Supervisor: Claire Miller

External Supervisor: Yvonne McElarney, AFBI, Northern Ireland

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Eutrophication is a major threat to freshwater systems and water quality across Europe continues to decline (European Environment Agency, 2018) despite management meaures implemented under the European Water Framework Directive (WFD) (European Commission, 2000).

Lough Neagh is a large lake in Northern Ireland, covering 383km^2, which is important for freshwater supplies to Northern Ireland and commercial fishing. Therefore, the nutrient loading and water chemistry of the surrounding catchment draining into the Lough is particularly important. This project will investigate the nutrient loading and water chemistry for 8 major in-flowing rivers to Lough Neagh. The rivers have been sampled manually weekly by the Agri-food and Biosciences Institute (AFBI), in Northern Ireland and hence time series data are available since the mid-1980s. The focus of interest for this project is the identification of trends and patterns in nutrient dynamics and water chemistry for these surrounding rivers, and hence potential implications for the impacts on Lough Neagh.

Please Note: The data are confidential and hence will be covered by a data licence. This will state that the data are only to be used by the student for this project and cannot be transferred to another party.

Key Questions of Interest

1. What are the changes in the loading of nutrients from the sub-catchments?
2. Is there evidence of trends and seasonal patterns for water chemistry in each of the rivers?
3. How do nutrient loading and water chemistry patterns differ for sub-catchments?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

A small amount of programming may be required for data manipulation

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Data transformations
Generalised linear models
Flexible regression - (Generalised) additive models
Time series – autocorrelation/partial autocorrelation functions

Project 49: Modelling biological systems with nonparametric models

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Biological systems, such as prey-predator models, often exhibit complex non-linear relationships that parametric statistical models struggle to describe. Nonparametric models are more suited to modelling such systems as their structure provides a more flexible modelling framework.

In this project, you will investigate and compare nonparametric statistical modelling methods on datasets obtained from biological dynamic systems. In particular, you will learn and fit the Reproducing Kernel Hilbert Space (RKHS) regression model and compare it with some of the more well-known nonparametric model such as smoothing splines.

This project will provide you with excellent training experience in advanced machine learning algorithms. Some R packages are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself.

Key Questions of Interest

- How does the RKHS model perform in comparison with the more commonly used parametric models?
- How does the RKHS model perform in comparison with the other nonparametric models?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The project will require you to simulate data that mimics biological systems in R. R will also be used to build nonparametric models.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models, Flexible Regression

Desirable: Kernel methods.

Project 50: Parameter inference in dynamical systems with nonparametric models

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Many processes in science and engineering can be described by dynamical systems based on nonlinear ordinary differential equations (ODEs). Often ODE parameters are unknown and not directly measurable. Since non-linear ODEs typically have no closed form solution, standard iterative inference procedures require a computationally expensive numerical integration of the ODEs every time the parameters are adapted, which in practice restricts statistical inference to rather small systems. To overcome this computational bottleneck, approximate methods based on gradient matching have recently gained much attention.

In this project, you will investigate a nonparametric statistical modelling method to estimate the parameters of the nonlinear dynamical system. The idea is to circumvent the numerical integration step by using a surrogate cost function that quantifies the discrepancy between the derivatives obtained from a smooth interpolant to the data and the derivatives predicted by the ODEs. In particular, you will learn and fit the Reproducing Kernel Hilbert Space (RKHS) regression model.

Some R packages are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself.

Key Questions of Interest

- How to estimate the ODEs parameters by using the RKHS model?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The project will require you to simulate data that mimics biological systems in R. R will also be used to build nonparametric models.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models

Desirable: Kernel methods.

Project 51: Modelling Atmospheric CO_2 level with Gaussian Process regression models

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

The atmospheric CO_2 observations from the Mauna Loa Observatory, Hawaii, are classic real-world examples for comparing different regression models. This dataset encodes nonlinear trends and periodic behaviour. In this project, you will investigate and compare nonparametric statistical modelling methods on this dataset. In particular, you will learn and fit the Gaussian process (GP) regression model and compare it with some of the more well-known nonparametric model such as smoothing splines and parametric models like linear regression models. Gaussian process regression is a nonparametric and Bayesian approach to regression.

This project will provide you with excellent training experience in advanced machine learning algorithms. Some R packages are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself.

Key Questions of Interest

- How does the GP model perform in comparison with the more commonly used parametric models?
- How does the GP model perform in comparison with the other nonparametric models?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

A small amount of python or R coding are required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
Linear model

Desirable:
Gaussian Process

Project 52: Sparse Gaussian Processes using variational inference

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

The application of Gaussian process (GP) models is intractable for large datasets because the time complexity scales as O(n^3) and the storage as O(n^2) where n is the number of training examples. To overcome this limitation, many approximate or sparse methods have been developed. These methods construct an approximation based on a small set of m support or inducing variables that allow the reduction of the time complexity.

In this project, we will investigate a variational formulation for sparse approximations that jointly infers the inducing inputs and the kernel hyperparameters by maximizing a lower bound of the true log marginal likelihood. The key property of this formulation is that the inducing inputs are defined to be variational parameters which are selected by minimizing the Kullback-Leibler divergence between the variational distribution and the exact posterior distribution.

This project will provide you with excellent training experience in advanced machine learning algorithms. Some Python packages are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself.

Key Questions of Interest

- How does the sparse GP model perform in comparison with the standard GP regression models?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount of Python or R coding are required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
Linear model
Maximum Likelihood Estimation

Desirable:
Gaussian Process
Kullback-Leibler divergence

Project 53: Gaussian Process Classification using variational inference

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Gaussian processes (GPs) provide priors over functions that can be used for many machine learning tasks. In the regression setting, when the likelihood is Gaussian, inference can be performed in closed form using linear algebra. When the likelihood is non-Gaussian, such as in GP classification, the posterior and marginal likelihood must be approximated. To perform classification with the GP prior, the process is 'squashed' through a sigmoidal inverse-link function, and a Bernoulli likelihood conditions the data on the transformed function values.

In this project, we will investigate a variational formulation approximation which can provide variational bounds with non-Gaussian likelihoods. The key property of this formulation is that the inducing inputs are defined to be variational parameters which are selected by minimizing the Kullback-Leibler divergence between the variational distribution and the exact posterior distribution.

This project will provide you with excellent training experience in advanced machine learning algorithms. Some Python packages are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself.

Key Questions of Interest

- How does the GP classification model perform in comparison with other standard parametric models?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount of Python or R coding are required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
Linear model
Maximum Likelihood Estimation

Desirable:
Gaussian Process
Kullback-Leibler divergence

Project 54: Dimension reduction using Gaussian Processes

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Summarising high dimensional dataset with a low dimensional space (or embedding) is a standard approach in statistifal modelling. The structure of the high dimensional dataset can be explored by dimension reduction. In this project, we will apply some existing dimension reduction techniques to simulation datasets and some real datasets.

The standard dimension reduction approach like Principle Component Analysis (PCA) can be compared with some more advanced machine learning techniques such as Gaussian process latent variable models. Similar to the idea of unsupervised learning, the original high dimensional data can be represented by the low dimensional latent space which is modelled by PCA or Gaussian process latent variable model.

Some video lectures, research papers and software packages are available to help the student to initialise the project.

Key Questions of Interest

1. What methods are appropriate to investigate the underling structure of the simulated dataset?
2. What methods are appropriate to investigate the underling structure of the real dataset of handwritten digits?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate/Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount of Python or R coding are required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
PCA
Linear model

Desirable:
Gaussian Process
Gaussian Process Latent variable model

Project 55: Bayesian nonlinear dimension reduction for handwritten digits

Statistics Supervisor: Mu Niu

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Gaussian Processes (GP) offer a Bayesian nonparametric framework for inference of highly nonlinear latent functions from observed data. They have become very popular in machine learning. The application of GPs to unsupervised learning tasks is more involved. It can summarise high dimensional dataset with a low dimensional space. One approach to unsupervised learning with GPs is the Gaussian process latent variable model (GP-LVM). In this project, we will investigate a variational inference framework for training the Gaussian process latent variable model and thus performing Bayesian nonlinear dimensionality reduction.

The standard dimension reduction approach like Principle Component Analysis (PCA) can be compared with some more advanced machine learning techniques such as Bayesian Gaussian process latent variable model. We will apply some existing dimension reduction techniques to simulation datasets and real datasets of handwritten digits (0-9).

Some video lectures, research papers and software packages are available to help the student to initialise the project.

Key Questions of Interest

1. What methods are appropriate to investigate the underling structure of the simulated dataset?
2. What methods are appropriate to investigate the underling structure of the real dataset of handwritten digits?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Difficult

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Python or R coding are required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential:
PCA
Linear model
Maximum likelihood estimates

Desirable:
Gaussian Process
Gaussian Process Latent variable model

Project 56: Measles Susceptibility in Scotland

Statistics Supervisor: Gary Napier

Can be adapted to: Single / Combined / MSciWP

Brief Description of Project

The Scottish Childhood Immunisation Record System (SCIRS) holds the individual records of all childhood vaccinations in Scotland. These include measles, mumps, and rubella (MMR) vaccination uptake, which occurs when a child is 12-13 months old and again at 4-5 years of age. Vaccines have been a discussion point for many years with the anti-vaccination movement and recent outbreaks of measles across Europe and North America.

Data are available from the SCIRS database at the intermediate zones (IZs) level in Scotland, which are small geographical units containing, on average, 4000 residents between 1998 and 2014. The beginning of this time period was when Wakefield et al. (1998) linked the MMR vaccine with an increased risk of autism, with the media coverage surrounding the article resulting in vaccination rates dropping to around 80% in 2003 in parts of the United Kingdom (McIntyre and Leask, 2008). These reduced vaccination rates later resulted in large outbreaks of measles in the UK in 2013 (Pollock et al., 2014). The article by Wakefield et al. (1998) was partially retracted in 2004, before being discredited in 2010 after several epidemiological studies failed to find any association with an increased risk in autism (Elliman and Bedford, 2007).

References

Elliman, D. and Bedford, H. (2007). MMR: where are we now? *Archives of Disease in Childhood* **92**, 1055-1057.

McIntyre, P. and Leask, J. (2008). Improving uptake of MMR vaccine. *British Medical Journal* **336**, 729-739.

Pollock, K., Potts, A., Love, J., Steedman, N. and Donaghy, M. (2014). Measles in Scotland, 2013. *Scottish Medical Journal* **59**, 3-4.

Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., Malik, M., Berelowitz, M., Dhillon, A.P., Thomson, M.A., Harvey, P., Valentine, A., Davies, S.E., Walker-Smith, J.A. (1998). Illeal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive development disorder in children. *The Lancet* **351**, 637-641.

Key Questions of Interest

The main questions of interest that will be investigated in this project are:

1. Did all IZs in Scotland exhibit a change in measles susceptibility following the retraction of the Wakefield article or did some areas display no change?

2. Did the change, if any, in measles susceptibility occur in 2004 alongside the articles' retraction for all IZs or was it earlier or later for some regions of the country?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

The project will be done in R with appropriate visualisations of the data produced, and Generalised Linear Models (GLMs) and spatio-temporal models fitted.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Generalised Linear Models (GLM) [essential]

Linear Mixed Models (LMM)

Spatial Statistics

Project 57: Hospital admissions due to respiratory disease in Greater Glasgow and Clyde

Statistics Supervisor: Gary Napier

Can be adapted to: Single / Combined / MSciWP

Brief Description of Project

Respiratory disease is only second to cancer as the most common cause of death in Scotland (http://www.gov.scot/Topics/Statistics/Browse/Health/TrendMortalityRates). Here, we focus on Greater Glasgow and Clyde because Glasgow is one of the unhealthiest cities in Europe (Gray et al., 2012).

Data are available at intermediate zone (IZ) level, which are small geographical units containing, on average, 4000 residents between 2002 and 2011 for the Greater Glasgow and Clyde Health Board. The spatial pattern and temporal trends in the number of hospital admissions due to respiratory disease will be modelled to determine which areas in Greater Glasgow and Clyde have exhibited changes in risk over the 10-year period.

References

Gray, L., Merlo, J., Mindell, J., Hallqvist, J., Tafforeau, J., O'Reilly, D., Regidor, E., Naess, O., Kelleher, C., Helakorpi, S., Lange, C. *and others* (2012). International differences in self-reported health measures in 33 major metropolitan areas in Europe. *European Journal of Public Health* **22**, 40-47.

Key Questions of Interest

The main questions of interest that will be investigated in this project are:

1. Which areas exhibit an increase, a decrease, or no change in risk over the 10-year period?

2. How have these changes in risk impacted upon health inequalities?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

The project will be done in R with appropriate visualisations of the data produced, and Generalised Linear Models (GLMs) and spatio-temporal models fitted.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Generalised Linear Models (GLM) [essential]

Linear Mixed Models (LMM)

Spatial Statistics

Project 58: Socioeconomic differences in Scottish higher education

Statistics Supervisor: Colette Mair

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

All 19 of Scotland’s higher education institutions have committed to widening access to university with a focus on Scottish students from disadvantaged areas. In 2014, the Scottish government set a target to have a fifth of all students come from disadvantages areas by 2030.

The commitments include proposals for improving access through changing admissions criteria, linking more with colleges and establishing more bridging programmes from school. For instance, students from disadvantaged backgrounds will be judged not only on academic achievement but on a wider range of achievements and potential.

The Scottish Index of Multiple Deprivation is a regional identifier and is commonly used a proxy for attainment. It is an official tool to identify areas of multiple deprivation in Scotland ranging from the most deprives (quantile 1) to most affluential (quantile 5) areas.

the Universities and Colleges Admissions Service in the UK (UCAS) provides independent in-depth analysis and insight about who’s applying and being accepted into full-time undergraduate higher education every year. Using application data between 2006 and 2019, this project with quantify age, gender and socioeconomical differences in applications to Scottish higher education institutions and assess if and how these differences change over time.

Key Questions of Interest

1. Are there differences in application acceptances between gender, age and/or socioeconomic status?
2. Have these differences changed over time?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Analysing data using packages in R.
No simulation is required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models, Generalised Linear Models, Time Series.
Desirable: flexible regression, Data Analysis.

Project 59: Cocirculation of the flu and cold.

Statistics Supervisor: Colette Mair

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Flu and the common cold are both respiratory illnesses, but they are caused by different viruses (influenza and rhinovirus respectively, amongst others). Because these two types of illnesses have similar symptoms, it can be difficult to tell the difference between them based on symptoms alone. In general, flu is worse than the common cold, and symptoms are more intense. Colds are usually milder than flu. People with colds are more likely to have a runny or stuffy nose. Colds generally do not result in serious health problems, such as pneumonia, bacterial infections, or hospitalizations. Flu can have very serious associated complications in children and the elderly.

Influenza viruses and rhinoviruses are responsible for many acute respiratory viral infections in human populations and are detected as co-pathogens within hosts. Clinical and epidemiological studies suggest that rhinovirus and influenza virus interfere with each other's spread within a population. For example, during summer and autumn of 2009 many European countries experienced a pandemic in the influenza A (H1N1) virus and it is believed that rhinovirus may have hampered the development of the influenza pandemic. Rhinovirus epidemics typically occur after the start of school terms in autumn and spring and this may interfere with the spread of influenza during this period when warm and humid climate decreases the influenza spread by aerosol. Limited laboratory data support this hypothesis.

Using monthly count data from influenza and rhinovirus between 2005 and 2013 from Greater Glasgow and Clyde, we will investigate temporal associations between these two viruses.

Key Questions of Interest

1. What is the relationship between the prevalence of flu and the cold?
2. Are coinfection rates different to what we would expect from two independent viruses?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

Analysing data using packages in R.
No simulation is required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models, Generalised Linear Models, Time Series.
Desirable: flexible regression, Data Analysis.

Project 60: Can Google trends predict healthcare seeking behaviour.

Statistics Supervisor: Colette Mair

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Flu like symptoms can be caused by different viruses and it can be difficult to tell the difference between them based on symptoms alone. For example, the flu and cold have similar symptoms but flu symptoms are more intense. Colds are usually milder than flu. People with colds are more likely to have a runny or stuffy nose. Flu can have very serious associated health complications.

Google is a popular search engine for accessing health-related information. For each search term, Google assesses the term's relative popularity on a scale between 0 and 100 by dividing the number of searched by the total number of overall searches during the specified time. The use of Google Trends to assess search trends can act as an indicator of how health-related coverage affects online information health-seeking behaviours of individuals. Data from Google Trends were extracted to provide a summary of queried search terms, phrases and keywords each month between 2005 and 2013.

Using monthly count data from primary and secondary healthcare providers from Greater Glasgow and Clyde between 2005 and 2013, this project explores the relationship between google trends from Scotland in searches relating to flu like symptoms and the number of people seeking healthcare with flu like symptoms.

Key Questions of Interest

1. What is the relationship between Google trends from Scotland and health care seeking behaviour in Greater Glasgow and Clyde?
2. Is there a lag effect?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Analysing data using packages in R.
No simulation is required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models, Generalised Linear Models, Time Series.
Desirable: flexible regression, Data Analysis.

Project 61: World’s Toughest Sports

Statistics Supervisor: Colette Mair

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

In 2004, a group of eight sporting experts ranked sixty sports based on ten skills thought to determine athleticism: endurance; strength; power; speed; agility; flexibility; coordination; nerve; durability and aptitude. The experts included sports scientists, academics, experts in muscle and movement, athletes and sports journalists. Each expert scored each sport on the ten attributes on a scale of 1 to 100. Difficulty scores were estimated based on average values between experts and the sports were ranked based on difficulty scores. It was determined that boxing is the toughest sport with a difficulty score of 72.375. However, can the ten attributes of athleticism be used to cluster sports? Moreover, what is the relationship between athlete earnings, based on salary, winnings and endorsement, and the level of difficulty of a sport? Do athletes who complete in more difficult sports earn more?

Using data from athletes listed on the Forbes highest paid athletes 2020, this project will assess the importance of these ten attributes in explaining athlete earnings and assess if these high earners experienced a drop in earnings during the 2020 pandemic.

Key Questions of Interest

1. Can the ten listed attributes of athleticism be used to cluster sports?
2. Has the global pandemic in 2020 impacted the total earnings of some of the highest paid athletes?
3. Which factors explain variation in athlete earnings?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Analysing data using packages in R.
No simulation is required.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential: Linear Models, Generalised Linear Models, Multivariate Methods

Project 62: Statistics Anxiety in Introductory Statistics Courses (A)

Statistics Supervisors: Mitchum Bock

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The aim of this project is to examine the relationship between students' performance in introductory university statistics courses and their self-efficacy (Schwarzer & Jerusalem, 1995) and attitudes towards Statistics (Cruise et al, 1985; Hanna et al, 2008) **at the start of the course**.

Some students display high levels of anxiety towards the subject of Statistics and this can adversely affect their performance in statistics courses. It might be anticipated that students with high levels of self-efficacy would be better able to overcome subject anxiety than other students.

This project will analyse data obtained from a small study of statistics anxiety that was carried out in the School of Mathematics & Statistics in Session 2019-20. Validated instruments measuring statistics anxiety and general self-efficacy were administered to students in **three different Introductory Statistics courses**: one intended for specialists in the mathematical sciences, an optional course intended for specialists in other disciplines and a compulsory course for specialists in another field. Appropriate ethical approval and participant consent were obtained.

Key Questions of Interest

The aims of the project are:

- to code and tidy the data obtained from the questionnaires in the three classes and the grades (where available);
- to summarise and compare the prevalence of statistics anxiety between the three courses;
- to investigate the relationship between statistics anxiety and grades and the type of course;
- to investigate the effect of self-efficacy on that relationship.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***No***

If ‘Yes’, please specify what this might involve:

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential techniques:

- Data management and manipulation in **R**.
- Data summaries, tables and plots.
- Confidence intervals and hypothesis tests.
- General linear models.
- Generalised linear models.

Project 63: Statistics Anxiety in Introductory Statistics Courses (B)

Statistics Supervisors: Mitchum Bock

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The aim of this project is to examine the relationship between students' performance in an introductory university statistics course and their self-efficacy (Schwarzer & Jerusalem, 1995) and attitudes towards Statistics (Cruise et al, 1985; Hanna et al, 2008) **at the start and end of the course**.

Some students display high levels of anxiety towards the subject of Statistics and this can adversely affect their performance in statistics courses. It might be anticipated that students with a high levels of self-efficacy would be better able to overcome subject anxiety than other students.

This project will analyse data obtained from a small study of statistics anxiety that was carried out in the School of Mathematics & Statistics in Session 2019-20. Validated instruments measuring statistics anxiety and general self-efficacy were administered to students at the **start and the end** of **two different Introductory Statistics courses**: one intended for specialists in the mathematical sciences, the other a compulsory course for specialists in another field. Appropriate ethical approval and participant consent were obtained.

Key Questions of Interest

The aims of the project are:

- to code and tidy the data obtained from the questionnaires in the two classes and the grades;
- to summarise and compare the prevalence of statistics anxiety within and between the two courses;
- to investigate the relationship between statistics anxiety and grades and the type of course;
- to investigate the effect of self-efficacy on that relationship.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***No***

If 'Yes', please specify what this might involve:

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential techniques:

- Data management and manipulation in `R`.
- Data summaries, tables and plots.
- Confidence intervals and hypothesis tests.
- General linear models.
- Generalised linear models.

Project 64: Statistics Anxiety in Introductory Statistics Courses (C)

Statistics Supervisors: Mitchum Bock

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The aim of this project is to examine the relationship between students' performance in introductory university statistics courses and their self-efficacy (Schwarzer & Jerusalem, 1995) and attitudes towards Statistics (Cruise et al, 1985; Hanna et al, 2008) **at the start of the course**.

Some students display high levels of anxiety towards the subject of Statistics and this can adversely affect their performance in statistics courses. It might be anticipated that students with high levels of self-efficacy would be better able to overcome subject anxiety than other students.

This project will analyse data obtained from **two small studies** of statistics anxiety that were carried out in the School of Mathematics & Statistics in the **2014-15 and 2019-20 sessions**. Validated instruments measuring statistics anxiety and general self-efficacy were administered to students in **two different Introductory Statistics courses**: one intended for specialists in the mathematical sciences and the other an optional course intended for specialists in other disciplines. Appropriate ethical approval and participant consent were obtained.

Key Questions of Interest

The aims of the project are:

- to code and tidy the data obtained from the questionnaires in the two classes and the grades (where available);
- to summarise and compare the prevalence of statistics anxiety within and between the two courses;
- to investigate the relationship between statistics anxiety and grades and the type of course;
- to investigate the effect of self-efficacy on that relationship.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy/Moderate

Is any Programming/Simulation required? ***No***

If ‘Yes’, please specify what this might involve:

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential techniques:

- Data management and manipulation in `R`.
- Data summaries, tables and plots.
- Confidence intervals and hypothesis tests.
- General linear models.
- Generalised linear models.

Project 65: Investigating Air Temperature Patterns in US Cities

Statistics Supervisor: Ruth O'Donnell

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

As data collection technologies are improving so too is the quantity and quality of data that are available, and we now have large quantities of time-series data which have arisen from the natural environment which we can use to study changes in climate.

One way to gain a better understanding of our natural environment is to explore environmental variables, such as air temperature, in terms of their patterns over space and time. We can also explore the relevance of additional covariates on these patterns where data are available. Furthermore, as data of environmental interest are usually collected at multiple stations, it is also often of interest to investigate temporal coherence amongst stations via unsupervised learning methods, such as clustering, in order to identify groups of stations that share similar characteristics.

In this project you will apply appropriate statistical methods to describe patterns in average daily air temperature at 48 US cities using temperature data covering the time period between 1995 and May 2020 as well as several covariates which may be of interest. The data are publicly available from the Average Daily Temperature Archive of the University of Dayton and can be found at

http://academic.udayton.edu/kissock/http/Weather/default.htm.

The available data are contained in a csv file called **USCityTemp** which contains the information on date, city name and

- Temperature in Fahrenheit
- Latitude of the city;
- Longitude of the city;
- Altitude of the city;
- Whether the city is by the coast (coded as 1) or not (coded as 0);

Key Questions of Interest

- What are the key temporal patterns in average daily air temperature at each of the 48 cities across the US for which data are available?
- Is there any relationship between the air temperature at each of the cities and the geographical characteristics of the city? (i.e. location, altitude, proximity to the coast)
- Can we identify any clusters of common temperature patterns amongst the cities? If yes, what is the optimal number of clusters within needed to describe the underlying temporal temperature dynamics.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***No***

If ‘Yes’, please specify what this might involve:

While there will be elements of R programming required to fit appropriate models and produce relevant plots, there is no simulation or function development needed for this project.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Essential
Linear Models
Flexible Regression (fitting Generalised Additive Models)
Multivariate Methods (clustering methods and possibly for dimension reduction)

Desirable
Time Series (assessing autocorrelation)
Functional Data Analysis

Project 66: Water quality in the Lunan catchment

Statistics Supervisor: Marian Scott

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Environmental quality is something of considerable interest, and the various agencies in Scotland routinely monitor water quality using a number of different measures, including temperature, dissolved oxygen and conductivity. New monitoring equipment means that data are being generated very frequently (in this case every hour), so that it is important to consider the seasonality and patterns observed in these records. Data from a small river in Scotland, covering the period 2008 to 2015 are considered. With the new technology there may also be questions concerning data quality assurance, and the reliability of the data generated.

Key Questions of Interest

1) To evaluate the data quality assurance, including approaches to deal with missing data, detecting outliers and extremes
2) To investigate seasonality and trends in water quality

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***No***

If 'Yes', please specify what this might involve:

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Time series (essential), regression and flexible regression (essential)

Project 67: Abundances of seabirds

Statistics Supervisor: Marian Scott

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Annual surveys of seabirds are conducted by volunteers around the country, these surveys record the numbers of birds each year. The results from the surveys over the years are used to track changes in species, both declines and increases. Changes can occur for many reasons including climate change, changes to habitat and food supply to mention only a few. Synchronicity in change in different species can be important to detect, since this can provide evidence about underlying causes. This project will look at the seabird abundances in Scotland, England and Wales over 20+ years and more than 30 species.

Key Questions of Interest

There are 2 main questions of interest:

a) How to describe the trends in seabird abundances
b) To assess which species are showing similar patterns and whether related to spatial location

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***/No***

If ‘Yes’, please specify what this might involve:

There may be a need to do some preliminary data preparation.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Regression and flexible regression (essential), multivariate analysis (essential), time series,(preferable)

Project 68: Wind speed

Statistics Supervisor: Marian Scott

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Renewable energy is a very important part of energy strategy, especially wind. This project will examine wind speeds measured at a location in the UK, and also modelled wind speed. By comparing the two wind speeds, the value of the model and whether it can be used to predict wind speeds can be ascertained, Additional meteorological variables including temperature and humidity are considered as possible predictors of windspeed.

Key Questions of Interest

1) To evaluate the cyclical and seasonal patterns in measured and modelled wind speeds and extreme values.
2) To investigate the relationship between the measured and modelled wind speeds and to build a model allowing the prediction of measured wind speed.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***No***

If 'Yes', please specify what this might involve:

There may need to be some preliminary data preparation

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Time series (essential), regression and flexible regression (essential), environmental statistics (and extremes) (desirable)

Project 69: Bayesian prediction of the price of gold

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Stochastic differential equations (SDE) are widely used in modelling financial processes such as interest rates, stock and commodity prices. We will be using existing SDE models of volatile markets to describe the behaviour of the price of gold. The aim of this analysis is to predict credible intervals for the price of gold in the future. Such an analysis requires inferring model parameters that match current history of the price. Unfortunately, the likelihood imposed by the SDE models does not have a closed form, and therefore traditional inference methods are not applicable to this problem.

To tackle the problem of likelihood intractability, we will be using the Approximate Bayesian Computation (ABC) methods for approximate inference and prediction.

Key Questions of Interest

1. How can inference be performed for problems where likelihood is not available in closed form?
2. Can we predict future observations of the price of gold using historical data?
3. How to formulate informative priors for a problem with good intuitive understanding of the subject?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

Historical data is available online, from gold.org. The student will need to implement several Approximate Bayesian Computation algorithms to perform inference for the Black-Scholes SDE model, and draw future predictions from this model. If time allows, several alternative models can be considered.

Programming inference algorithms in R should be sufficient for this project.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

The students should be aware of Bayesian approach to inference and would benefit from previous experience with sampling for inference.

Bayesian Statistics and Advanced Bayesian Methods are two courses that would be helpful to address this project.

Project 70: Weather forecasting using Recursive Neural Networks

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Artificial Neural Networks have become a fashion in machine learning within the last decade. Proponents of this approach claim impressive capabilities of this type of models in classification and forecasting problems. This type of models is claimed to be very flexible for modelling complex behaviours.

In this project we will build several simple models using artificial neural networks and apply them to forecasting of time series data, applied particularly to a dataset of weather measurements. The prediction results will be compared to a naïve approach of saying that the weather tomorrow is going to be exactly the same as today, as well as to forecasts made using linear and ARMA models.

Recursive Neural Networks will be defined and fitted using Tensorflow framework.

Key Questions of Interest

1. How to define and fit artificial neural networks using Tensorflow?
2. Do neural network models provide any benefit over linear modelling?
3. How to setup a typical workflow for supervised machine learning?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Hard

Is any Programming/Simulation required? ***Yes***

A historic weather data set is available online. We will look at the time series in this data set, and investigate how to predict future observations using a number of alternative models.

For the neural network modelling, we will be using Tensorflow to define and fit models. This will require additional study to learn basics of Python programming and using Tensorflow.

Forecasting results will be compared using verification data.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Time series analysis is the core of this project. Benchmark models are the typical models considered in Time Series and Linear Models courses. Neural network modelling is extracurricular study, and will require the student to learn a little of Python programming. Students confident with R programming will find this part to be relatively easy.

Project 71: Bayesian analysis of crime statistics

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

Linear models are the most ubiquitous class of statistical models used in practice. In your courses these models were covered extensively using the classical approach. In this project, you will consider the Bayesian approach to inference using linear models and will consider the problem of variable selection using Lasso regularisation in Bayesian framework.

In particular we will be looking at a dataset of recorded violent crimes in 1994 different US neighbourhoods. We will establish what demographical characteristics impact the crime rate the most.

Key Questions of Interest

1. How to formulate a linear model in the Bayesian framework?
2. How to perform conjugate and non-conjugate inference for a Bayesian linear model?
3. How to perform variable selection using Bayesian Lasso?
4. What are the most important factors to explain variation in crime rates among different neighbourhoods?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

Demographic data for different US neighbourhoods is available online. One of the variables in this data set is the rate of violent crime per 100K population. We will define a linear model to perform regression on these data, and perform Bayesian inference of model parameters.

Inference will involve Gibbs and Metropolis-Hastings sampling.

Bayesian Lasso will be implemented using regularising priors on regression coefficients.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

This is a typical analysis task using methods covered in Bayesian Statistics and Advanced Bayesian Methods courses.

Project 72: What makes a good wine?

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP (delete only if necessary)

Brief Description of Project

We will consider an extensive dataset containing descriptions of 4898 wines. The response variable is a score for the quality of a particular wine. Other 11 variables include measurements of acidity, residual sugar, chlorides, sulphates, density, etc. To investigate interaction between explanatory variables, you may want to introduce latent explanatory variables, such as pairwise products of the above variables, or their ratios.

We will use a Bayesian formulation for linear regression and perform Bayesian Lasso for variable selection to decide what are the most important factors that make a good wine.

Key Questions of Interest

1. How to formulate a linear model in the Bayesian framework?
2. How to perform conjugate and non-conjugate inference for a Bayesian linear model?
3. How to perform variable selection using Bayesian Lasso?
4. What are the most important factors to explain variation in wine scores?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

Wine quality dataset is available online. We will define a linear model to perform regression on these data, and perform Bayesian inference of model parameters.

Inference will involve Gibbs and Metropolis-Hastings sampling.

Bayesian Lasso will be implemented using regularising priors on regression coefficients.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

This is a typical analysis task using methods covered in Bayesian Statistics and Advanced Bayesian Methods courses.

Project 73: How to make the strongest concrete ever?

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP   (delete only if necessary)

Brief Description of Project

We will consider a dataset containing descriptions of 103 concrete mixtures. The response variable is a measurement of compressive strength after 28 days curing tme. Other 9 variables are proportions of components in the mixture. To investigate interaction between explanatory variables, you may want to introduce latent explanatory variables, such as pairwise products of the above variables, or their ratios.

We will use a Bayesian formulation for linear regression and perform Bayesian Lasso for variable selection to decide what are the most important factors to making strong concrete.

Key Questions of Interest

1. How to formulate a linear model in the Bayesian framework?
2. How to perform conjugate and non-conjugate inference for a Bayesian linear model?
3. How to perform variable selection using Bayesian Lasso?
4. What are the most important factors to explain variation in concrete strength?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

Concrete strength dataset is available online. We will define a linear model to perform regression on these data and perform Bayesian inference of model parameters.

Inference will involve Gibbs and Metropolis-Hastings sampling.

Bayesian Lasso will be implemented using regularising priors on regression coefficients.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

This is a typical analysis task using methods covered in Bayesian Statistics and Advanced Bayesian Methods courses.

Project 74: Bayesian Changepoint Detection

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined / MSciWP

Brief Description of Project

In this project you will perform Bayesian analysis of a coal mining disasters data set.

Having observations of events coming from a Poisson process, you will use Bayesian framework of dealing with uncertainty to decide if the rate of the underlying process stays constant or whether it changes over time.

You will implement an inference method that will detect likely positions of changepoints, where the rate of the Poisson process has likely changed.

You will perform some form of Bayesian hypotheses testing to decide how many change points are required to explain the observed data set.

Key Questions of Interest

How to make decisions in case of uncertainty.

For your project you need to

1. Implement one of the inference approaches (can be direct sampling, Metropolis-Hastings, sequential importance sampling, or anything else) to calculate the posteriors of changepoint locations and corresponding disaster rates using a semi-conjugate prior.
2. Implement sampling from the posterior predictive distribution to see how well the model explains your observations.
3. Implement marginal likelihood evaluation procedures to decide on the number of changepoints most suitable to explain data.
4. Consider at least a few simpler cases, where the number of combinations for changepoints is not too dramatic, and show your inference and prediction results.
5. Write a report describing your work, and discussing your observations and experience.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

Direct simulations will certainly work (as they were tried before) but will not handle more complex models (over about 5 changepoints)

Metropolis, or sequential importance sampling will likely handle any model for this problem.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

1. One of the sampling approaches, be it direct sampling, Metropolis-Hastings, or sequential importance sampling.
2. Changepoint models

Project 75: Bayesian Navigation

Statistics Supervisor: Vladislav Vyshemirsky

External Supervisor (if any):

Can be adapted to: Single / Combined / MSciWP

Brief Description of Project

In this project you will build a fully autonomous artificial intelligence algorithm for estimating your location on a map using the position resection method.

Having measured azimuth to several reference points on a map, you will perform inference of your current position using Bayesian framework of dealing with uncertainty.

At the second stage of the project, you would assume that your position is on a roadway, and incorporate that information in your inference by selecting an appropriate prior.

Key Questions of Interest

How to make decisions in case of uncertainty.

For your project you need to

1. Collect data. You can pick three reference points on an online map, and get their coordinates. Then you can either calculate bearing angles to your location to create a simulated data set, or take a compass and perform real measurements. Extra brownie points for actually measuring bearings.
2. Implement one of the inference approaches to this problem (can be Metropolis-Hastings, sequential importance sampling, or anything else), and calculate the posteriors of your location using a wide uniform prior, and then assuming proximity to a road.
3. Write a report describing your work, and discussing your observations and experience.

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Easy

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

Some calculations in R will suffice. Hints will be provided for dealing with map data.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

1. One of the sampling approaches, be it Metropolis-Hastings, sequential importance sampling, or Hamiltonian Monte-Carlo.
2. Dealing with geolocation data

Project 76: Are fish more stressed in connected freshwaters?

Statistics Supervisor: Craig Wilkie

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Hydroscape (https://hydroscapeblog.wordpress.com/) is a four-year project funded by the UK Natural Environment Research Council (NERC), which aims to determine how stressors and connectivity interact to influence biodiversity and ecosystem function in freshwaters across Britain. While stressors such as nutrient pollution and climate change drive ecological degradation, connectivity between freshwater habitats is a major force behind both dispersal of stressors and biodiversity. Currently, the implication for freshwaters of future changes in stressor intensity and in connectivity levels across Britain are poorly understood. Hydroscape will significantly improve this understanding and therefore inform the work of organisations engaged in waterbody restoration, biological conservation, the control of invasive species and diseases of wildlife and humans, at the international, national and local level.

There are two related projects here. (See the project titled “Are birds more stressed in connected freshwaters?”.)

This project will investigate the species richness of fish across bodies of freshwater in different localised areas of Great Britain, over an aggregated period of 2002-2012. In particular, it is of interest to investigate how the relationships between stressors (e.g. % agricultural land, % urban land) and species richness change depending on landscape characteristics and connectivity of waterbodies.

Key Questions of Interest

1) What are the relationships between species distribution and stressors such as % of agricultural land and % of urban land in the surrounding catchment?
2) What are the relationships between species distribution and surrounding catchment characteristics such as number of lakes in the catchment?
3) Does the relationship between species distribution and stressors change depending on freshwater connectivity in the catchment?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If ‘Yes’, please specify what this might involve:

A small amount will be required for data manipulation.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Normal linear models (essential)
Generalised linear models (essential)
Flexible regression - Generalised additive models (desirable)
Knowledge of Environmental Statistics might be useful

Project 77: Are birds more stressed in connected freshwaters?

Statistics Supervisor: Craig Wilkie

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

Hydroscape (https://hydroscapeblog.wordpress.com/) is a four-year project funded by the UK Natural Environment Research Council (NERC), which aims to determine how stressors and connectivity interact to influence biodiversity and ecosystem function in freshwaters across Britain. While stressors such as nutrient pollution and climate change drive ecological degradation, connectivity between freshwater habitats is a major force behind both dispersal of stressors and biodiversity. Currently, the implication for freshwaters of future changes in stressor intensity and in connectivity levels across Britain are poorly understood. Hydroscape will significantly improve this understanding and therefore inform the work of organisations engaged in waterbody restoration, biological conservation, the control of invasive species and diseases of wildlife and humans, at the international, national and local level.

There are two related projects here. (See the project titled "Are fish more stressed in connected freshwaters?".)

This project will investigate the species richness of birds across bodies of freshwater in different localised areas of Great Britain, over an aggregated period of 2002-2012. In particular, it is of interest to investigate how the relationships between stressors (e.g. % agricultural land, % urban land) and species richness change depending on landscape characteristics and connectivity of waterbodies.

Key Questions of Interest

1) What are the relationships between species distribution and stressors such as % of agricultural land and % of urban land in the surrounding catchment?
2) What are the relationships between species distribution and surrounding catchment characteristics such as number of lakes in the catchment?
3) Does the relationship between species distribution and stressors change depending on freshwater connectivity in the catchment?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount will be required for data manipulation.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Normal linear models (essential)
Generalised linear models (essential)
Flexible regression - Generalised additive models (desirable)
Knowledge of Environmental Statistics might be useful

Project 78: How does river level and flow affect water quality?

Statistics Supervisor: Craig Wilkie

External Supervisor (if any):

Can be adapted to: Single / Combined/ MSciWP

Brief Description of Project

The Ramganga Water Data Fusion project (https://surajitstat.github.io/ramganga/) is a 2-year project investigating water quality in the Ramganga river catchment in India. The Ramganga river is a vital source for drinking water and irrigation, but it is highly polluted by industries and settlements in its catchment and it is necessary to understand its behaviour.

Water quality parameters are measured at various sampling stations in the river, throughout the year. The levels of various chemicals and pollutants in the river are affected by inflow from human settlements and industries, but their recorded concentrations are also affected by the volume of water flowing in the river at any time point.

This project will investigate how various water quality parameters are affected by river level, flow and rainfall over the surrounding catchment, using monthly historical data available from 1967 to 2012 for up to 4 sites in the river.

Key Questions of Interest

1) What are the patterns over time in river level, flow, rainfall and water quality parameters at each sampling station?
2) How do these parameters change in value over space?
3) How do river level, flow and rainfall relate to water quality at each location?

Analysis Summary

What level of difficulty do you think the project will have for the typical student?

Moderate

Is any Programming/Simulation required? ***Yes***

If 'Yes', please specify what this might involve:

A small amount will be required for data manipulation.

Please specify the statistical techniques which the project is **likely to require**, and any that are **essential** (as combined students may not have covered them, since they have options):

Normal linear models (essential)
Flexible regression - Generalised additive models (desirable)
Functional data analysis (desirable)
Knowledge of Environmental Statistics might be useful