University of Glasgow

MSc Projects in Statistics 2020/21

Contents

1 Spatio-temporal prediction of GP prescription rates (1)

Project level: Moderate

1.1 Overall project description

An important problem in data science is predictive modelling, where the aim is to use a set of collected data to predict a future or unmeasured observation. It is used in numerous application areas, ranging from predicting the betting odds of winning a sporting event, predicting the demand for health care services on a daily basis, and predicting which products a consumer is likely to buy to allow targeted online advertising via social media. However, predicting the unknown is a much more challening task then modelling observed data, and the assessment of predictive ability is different from simply looking at model fit statistics like AIC. This project will focus on how predictable unknown observations are, and the key challenges are to build and compare a range of models for their predictive ability, and quantify the accuracy with which the outcomes can be successfully predicted.

1.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Primary (non-hospitalised) care is delivered by groups of doctors located within GP (General practice) surgeries, who prescribe medicines called a prescription to people who are ill. There are 2 outcome variables to be separately modelled and predicted in this study, which relate to the rate of prescriptions that prevent (Corticosteroids) and relieve (short-acting $\beta-2$ agonists) the symptoms of respirtory conditions, such as asthma or chronic obstructive pulmonary disease. Surgeries with high rates prescribe more medications (after adjusting for the number and age/sex profile of their patient populations) than surgeries with lower rates, and a rate of 1 corresponds to an average rate for Scotland, while a rate of 1.2 corresponds to a 20% increased rate. The data are stored in `Prediction project.csv` and contain the following columns.

- `code` - A unique code for each GP surgery.
- `month` - The month the data relate to.
- `easting, northing` - The geographical coordinates of the GP surgery.
- `RATEprev` - The rate of prescriptions that prevent the symptoms of respirtory disease.
- `RATEreli` - The rate of prescriptions that relieve the symptoms of respirtory disease.
- `board` - The health board (part of Scotland) that the GP surgery is in.
- `dispensing` - Whether the GP surgery dispenses its own prescriptions.
- `perc_white` - The percentage of the patient population who are white.

- `price_med` - The average property price around the GP surgery.
- `pm10, pm25` - Measures of 2 different air pollutants.

Question(s) of interest

The main questions of interest are:

- What is the best model for predicting the rate of respiratory medications, and how predictiable are the rates?
- Which of reliever and preventer medications are most predictable?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics or flexible regression.

2 Spatio-temporal clustering and trend estimation (2)

Project level: Moderate

2.1 Overall project description

Many different data sets are collected at fixed spatial and temporal scales, such as house prices and ill health measures, and interest lies in identifying the spatio-temporal dynamics in these data. For example, one might be interested in identifying the region-wide overall temporal trend, or identifying if there are any spatial clusters in the data and how consistent these clusters might be over time. The data for each project form a regular array of spatio-temporal observations, where data are collected on the same set of spatial units for all time periods.

2.2 Individual project details

How many individual projects are available in this area: 2.

2.3 Property prices

Data available
Data are available on the average (median) selling price of properties that have sold in each year between 1993 and 2013 for each intermediate zone (IZ) in Scotland. Intermediate zones are small spatial areas created for the distribution of small-area statistics (see https://statistics.gov.scot/home), and the average population of each IZ is around 4,000 people. The data are stored in `spacetime project 1.csv` and contain the following columns.

- `IZ` - A unique code for each IZ area.
- `Area` - The name of the IZ area.
- `Y1993,...,Y2013` - The median property price of all properties sold in that year and IZ.

Question(s) of interest
The main questions of interest are:

- What are the main spatio-temporal dynamics of property prices, such as: (a) Scotland wide temporal trend; (b) changes in spatial variation in prices over time; (c) highest, lowest and biggest change areas between 1993 and 2013.
- How many clusters of IZs with similar property prices are there, and how consistent are these clusters over time?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate methods.

- One (or more) of Time series / Spatial statistics / Functional data analysis depending on which way you want to take the project.

2.4 Respiratory prescription rates

Data available

Primary (non-hospitalised) care is delivered by groups of doctors located within GP (General practice) surgeries, who prescribe medicines called a prescription to people who are ill. There are 2 outcome variables to be separately modelled and predicted in this study, which relate to the rate of prescriptions that prevent (Corticosteroids) and relieve (short-acting $\beta - 2$ agonists) the symptoms of respirtory conditions, such as asthma or chronic obstructive pulmonary disease. Surgeries with high rates prescribe more medications (after adjusting for the number and age/sex profile of their patient populations) than surgeries with lower rates, and a rate of 1 corresponds to an average rate for Scotland, while a rate of 1.2 corresponds to a 20% increased rate. The data are at a monthly resolution between October 2015 and August 2016 and relate to each GP surgery in Scotland. The data are stored in `spacetime project 2.csv` and contain the following columns.

- `code` - A unique code for each GP surgery.
- `easting, northing` - The geographical coordinates of the GP surgery.
- `prev.oct15,...,prev.aug16` - The rate of prescriptions that prevent the symptoms of respirtory disease in the months given.
- `reli.oct15,...,reli.aug16` - The rate of prescriptions that relieve the symptoms of respirtory disease in the months given.

Question(s) of interest

The main questions of interest are:

- What are the main spatio-temporal dynamics of respiratory prescribing, such as: (a) Scotland wide temporal trend; and (b) changes in spatial variation in rates over time.
- How many clusters of GPs with similar respiratory prescription rates are there, and how consistent are these clusters over time?
- How similar are the clusters and spatio-temporal dynamics between prescriptions that prevent and relieve the symptoms of respiratory illness.

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate methods.
- One (or more) of Time series / Spatial statistics / Flexible regression depending on which way you want to take the project.

3 Quantifying the impact of air pollution on human health (3)

Project level: Moderate

3.1 Overall project description

The health impact of exposure to air pollution is thought to reduce average life expectancy by six months, with an estimated equivalent health cost of 19 billion each year (from DEFRA). These effects have been estimated using statistical models, which quantify the impact on human health of exposure in both the short and the long term. However, the estimation of such effects is challenging, because individual level measures of health and pollution exposure are not available. Therefore, the majority of studies are conducted at the population level, and the resulting inference can only be made about the effects of pollution on overall population health. In this project you will utilise data on air pollution concentrations, disease incidence and other confounders, to estimate the health impact of air pollution.

3.2 Individual project details

How many individual projects are available in this area: 3.

3.3 Spatial analysis of hospitalisations

Data available

Data are available on air pollution, health and confounders in 2015 - 2016 for each intermediate zone (IZ) in Scotland, which are small spatial areas created for the distribution of small-area statistics. For details see https://statistics.gov.scot/home, and the average population of each IZ is around 4,000 people. The data are stored in `Air pollution and health project 1.csv` and contain the following columns.

- `IZ` - A unique code for each IZ area.
- `name` - The name of the IZ area.
- `Y_hosp_resp` - The number of respiratory hospitalisations in each IZ in 2015-2016.
- `E_hosp_resp` - The expected number of respiratory hospitalisations in each IZ in 2015-2016 based on population size and demographic structure.
- `employment, income, crime, housing, health, education, access` - Measures of socio-economic deprivation (poverty) for each IZ, which make up the Scottish Index of Multiple Deprivation.
- `no2, nox, pm10, pm25` - Measures of 4 different air pollutants.

Question(s) of interest

The main questions of interest are:

- What is the effect of each air pollutant on the risk of respiratory hospitalisation in Scotland?
- Which areas have the highest and lowest risk of disease?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

3.4 Spatio-temporal analysis of GP prescriptions

Data available
Data are available on air pollution, health and confounders at a monthly resolution for 11 months spread over 2015 - 2016 for each GP (doctors) surgery in Scotland, which provide primary (non-hospitalised) care for people who are unwell. The health outcome is the number of prescriptions for respiratory medicines that prevent (Corticosteroids) and relieve (short-acting $\beta - 2$ agonists) the symptoms of respirtory conditions, such as asthma or chronic obstructive pulmonary disease. The data are stored in `Air pollution and health project 2.csv` and contain the following columns.

- `code` - A unique code for each GP surgery.
- `month` - The month the data relate to.
- `december` - An indicator variable for December.
- `Yreli` - The number of prescriptions that relieve the symptoms of respirtory disease.
- `Yprev` - The number of prescriptions that prevent the symptoms of respirtory disease.
- `Ereli` - The expected number of prescriptions that relieve the symptoms of respirtory disease based on the GP population size and demographic structure.
- `Eprev` - The expected number of prescriptions that prevent the symptoms of respirtory disease based on the GP population size and demographic structure.
- `board` - The health board (part of Scotland) that the GP surgery is in.
- `easting, northing` - The geographical coordinates of the GP surgery.
- `dispensing` - Whether the GP surgery dispenses its own prescriptions.
- `perc_white` - The percentage of the patient population who are white.
- `price_med` - The average property price around the GP surgery.
- `pm10, pm25` - Measures of 2 different air pollutants.
- `urban` - How urban the area is that the GP surgery is located within.

Question(s) of interest
The main questions of interest are:

- What is the effect of each air pollutant on the rate of respiratory medications prescribed by GPs?
- Which health boards have the highest and lowest rates of respiratory prescriptions?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics or flexible regression.

3.5 Temporal analysis of deaths

Data available
Data are available on air pollution, health and confounders at a daily temporal resolution for 14 years from 1987 to 2000 for Chicago, USA. The data are stored in `Air pollution and health project 3.csv` and contain the following columns.

- `death` - Number of deaths (excluding accidents) in Chicago on that day.
- `temperature` - The average temperature.
- `day` - The day of the study, where 1 represents 1st January 1987, 2 is the 2nd January 1987 and so on.
- `no2, o3, pm10, so2` - Measures of 4 different air pollutants.

Question(s) of interest
The main questions of interest are:

- What is the effect of each air pollutant on the risk of non-accidental mortality in Chicago?
- What is the temporal trend in disease incidence?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Time series or flexible regression.

4 Modelling the relationship between deprivation and disease risk in Scotland (3)

Project level: Moderate

4.1 Overall project description

Scotland is often regarded as the 'sick man of Europe' as a result of the country's poor health compared to other European countries. This poor health has often been linked to social and economic inequality within the country. An NHS Scotland report from 2016 suggested that people living in poorer areas of Scotland were at a higher risk of disease than those living in wealthier areas. Modelling disease risk at the individual level can be challenging; health data on individuals is generally not available due to confidentiality concerns. Instead, the modelling tends to be carried out at the regional level, using administrative data provided by the Scottish government and regional health boards. In this project, you will model the relationship between disease risk and of a number of measures of deprivation at the population level.

4.2 Individual project details

How many individual projects are available in this area: 3.

4.3 Cancer

Data available

The dataset contains counts of the number of people admitted to hospital with particular diseases in 2012 each Intermediate Zone (IZ) in Scotland. These IZs are small regions created by Scottish government for administrative reasons; there are 1235 such regions, and the median population is just over 4000. Additionally, the dataset contains a number of covariates which can be considered proxy measures of socio-economic deprivation.

This data was obtained from the Scottish government via https://statistics.gov.scot/home. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `ScotlandData.csv` and contain the following columns.

- `id` - A unique code for each Intermediate Zone (IZ).
- `name` - The name of the IZ.
- `cancer, CHD, resp` - The number of hospital admissions related to cancer, coronary heart disease (CHD) and respiratory disease in each IZ in 2012.
- `population` - The estimated population of each IZ in 2012.
- `benefits, education, house_price, income, smoking` - Measures of socio-economic deprivation in each IZ - these include the percentage of people on benefits,

the percentage of people with 5 or more Standard Grade school qualifications, median house price, median weekly household income and percentage of pregnant mothers who smoked.

Question(s) of interest
The main questions of interest are:

- What is the relationship between **cancer** risk and these measures of socio-economic deprivation?
- Which areas have the highest and lowest risk of disease?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

4.4 Coronary Heart Disease (CHD)

Data available
The dataset contains counts of the number of people admitted to hospital with particular diseases in 2012 each Intermediate Zone (IZ) in Scotland. These IZs are small regions created by Scottish government for administrative reasons; there are 1235 such regions, and the median population is just over 4000. Additionally, the dataset contains a number of covariates which can be considered proxy measures of socio-economic deprivation.

This data was obtained from the Scottish government via https://statistics.gov.scot/home. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `ScotlandData.csv` and contain the following columns.

- `id` - A unique code for each Intermediate Zone (IZ).
- `name` - The name of the IZ.
- `cancer, CHD, resp` - The number of hospital admissions related to cancer, coronary heart disease (CHD) and respiratory disease in each IZ in 2012.
- `population` - The estimated population of each IZ in 2012.
- `benefits, education, house_price, income, smoking` - Measures of socio-economic deprivation in each IZ - these include the percentage of people on benefits, the percentage of people with 5 or more Standard Grade school qualifications, median house price, median weekly household income and percentage of pregnant mothers who smoked.

Question(s) of interest
The main questions of interest are:

- What is the relationship between **CHD** risk and these measures of socio-economic deprivation?
- Which areas have the highest and lowest risk of disease?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

4.5 Respiratory Disease

Data available
The dataset contains counts of the number of people admitted to hospital with particular diseases in 2012 each Intermediate Zone (IZ) in Scotland. These IZs are small regions created by Scottish government for administrative reasons; there are 1235 such regions, and the median population is just over 4000. Additionally, the dataset contains a number of covariates which can be considered proxy measures of socio-economic deprivation.

This data was obtained from the Scottish government via https://statistics.gov.scot/home. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `ScotlandData.csv` and contain the following columns.

- `id` - A unique code for each Intermediate Zone (IZ).
- `name` - The name of the IZ.
- `cancer, CHD, resp` - The number of hospital admissions related to cancer, coronary heart disease (CHD) and respiratory disease in each IZ in 2012.
- `population` - The estimated population of each IZ in 2012.
- `benefits, education, house_price, income, smoking` - Measures of socio-economic deprivation in each IZ. These include the percentage of people on benefits, the percentage of people with 5 or more Standard Grade school qualifications, median house price, median weekly household income and percentage of pregnant mothers who smoked.

Question(s) of interest
The main questions of interest are:

- What is the relationship between **respiratory disease** risk and these measures of socio-economic deprivation?
- Which areas have the highest and lowest risk of disease?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

5 Identifying contributory factors for disease mortality in England (2)

Project level: Moderate

5.1 Overall project description

Disease mortality tends to vary geographically as a result of the social, economic and environmental factors associated with different regions. Public Health England that substantial health inequalities are present in England; the gap in life expectancy between the most and least deprived parts of the country are 9.3 years for males and 7.3 years for females. In this project, you will construct a statistical model to identify some of the factors which might contribute to this disease inequality across England. Modelling disease risk at the individual level can be challenging; health data on individuals is generally not available due to confidentiality concerns. Instead, the modelling tends to be carried out at the regional level, using administrative data provided by the UK government and regional health boards.

5.2 Individual project details

How many individual projects are available in this area: 2.

5.3 Chronic Obstructive Pulmonary Disease (COPD)

Data available

The dataset contains counts of the number of deaths from chronic obstructive pulmonary disease and cancer in each Local Authority District (LAD) in England. The country is divided into 324 LADs for the purposes of local government, based on a combination of historic significance and administrative convenience. The dataset also contains a number of social, economic and environmental covariates relating to each region.

This data was obtained from the UK government via https://data.gov.uk. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `EnglandData.csv` and contain the following columns.

- `name` - The name of the Local Authority District (LAD).
- `id_short, id_long` - A pair of unique codes for each LAD.
- `COPD, cancer` - The number of deaths from Chronic Obstructive Pulmonary Disease (COPD) and cancer in each LAD in the most recently available year of data.
- `population` - The estimated population of each LAD in 2012.
- `poverty, education, unemployment, crime, pollution` - Social, economic and environmental measures for each region - these include the percentage of children in

poverty, the percentage of people with 5 or more GCSE school qualifications, the percentage of working age adults who are unemployed, the number of violent crimes per 1000 people and net CO2 pollution levels.

Question(s) of interest
The main questions of interest are:

- Which social, economic or environmental factors may contribute to **COPD** mortality in England?
- Which areas have the highest and lowest risk of disease?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

5.4 Cancer

Data available
The dataset contains counts of the number of deaths from chronic obstructive pulmonary disease and cancer in each Local Authority District (LAD) in England. The country is divided into 324 LADs for the purposes of local government, based on a combination of historic significance and administrative convenience. The dataset also contains a number of social, economic and environmental covariates relating to each region.

This data was obtained from the UK government via https://data.gov.uk. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `EnglandData.csv` and contain the following columns.

- `name` - The name of the Local Authority District (LAD).
- `id_short, id_long` - A pair of unique codes for each LAD.
- `COPD, cancer` - The number of deaths from Chronic Obstructive Pulmonary Disease (COPD) and cancer in each LAD in the most recently available year of data.
- `population` - The estimated population of each LAD in 2012.
- `poverty, education, unemployment, crime, pollution` - Social, economic and environmental measures for each region - these include the percentage of children in poverty, the percentage of people with 5 or more GCSE school qualifications, the percentage of working age adults who are unemployed, the number of violent crimes per 1000 people and net CO2 pollution levels.

Question(s) of interest
The main questions of interest are:

- Which social, economic or environmental factors may contribute to **cancer** mortality in England?
- Which areas have the highest and lowest risk of disease?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

6 Count data regression models (4)

Project level: Easy

6.1 Overall project description

Observational and epidemiological studies often give rise to count data, representing the number of occurrences of an event within some region in space or period of time, e.g., number of goals in a football match, number of emergency hospital admissions during a night shift, etc. A standard approach to modelling count data is Poisson regression: the counts are assumed to be independent Poisson random variables, with means determined, through a link function (usually the log), by a linear regression on available covariates. The Poisson model entails that the mean and variance are equal (equidispersion).

However, count data frequently exhibit underdispersion or, especially, overdispersion (these are often just symptoms of model misspecification, e.g. omission of important covariates, presence of outliers, lack of independence, inadequate link function).

The following projects deal with count data in which students need to use alternative models, compared to the Poisson model, to fit the data.

6.2 Individual project details

How many individual projects are available in this area: 4.

6.3 Number of children

Data available

This project uses data from Winkelmann (1995) on the number of births given by a cohort of women in Germany. The data consist of 1243 women over 44 in 1985 (and are located on the `fertility` data set which is located in the `Countr` library). The explanatory variables that were used can be seen below.

A data frame with 9 variables (5 factors, 4 integers) and 1243 observations:

- *children* integer; response variable: number of children per woman (integer),
- *german* factor; is the mother German? (yes or no),
- *years_school* integer; education measured as years of schooling,
- *voc_train* factor; vocational training ? (yes or no),
- *university* factor; university education ? (yes or no),
- *religion* factor; mother's religion: Catholic, Protestant, Muslim or Others (reference),
- *year_birth* integer; year of birth (last 2 digits),

- *rural* factor; rural (yes or no ?),
- *age_marriage* integer; age at marriage,

Question(s) of interest
The main questions of interest are:

- Is there a relationship between the previous variables and the number of children a woman has?
- Which model would you use to fit this data and why? (i.e. what is your main purpose of fitting this model? inference or prediction?)

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.

6.4 Heavy metal music and negative self-perception

Data available
Download the data set `project_heavy_metal.sav`. Install the package `memisc` and load the data in R using `Data <- as.data.set(spss.system.file("project_heavy_metal.sav"))`.

The data set is comprised of 121 rows and 15 columns.

- *age* denotes the age of a person,
- *age_group* denotes in which age group the person belongs to. The value 1 refers to $14-16$ years old while 2 refers to $16-19$ years old,
- *drug_use* refers to how many times the person used drug substances during the last year,
- *father_negligence* and *mother_negligence* are scales in which a high score is associated with a perception of cold and rejecting family relationships,
- *gender* shows the gender of the person,
- *isolation* corresponds to a subjective perception of lack of support,
- *marital_status* shows the marital status of the person's parents. The two possible values are *together* or *separated/divorced*,
- *meaninglessness* describes youths that may doubt the relevance of school in attaining future employment,
- *metal* describes how much the person listens to metal music,
- *normlessness* is defined as a belief that socially disapproved behaviours may be used to achieve certain goals,
- *self_estrangement* refers to persons who have a negative self-perception and who are overwhelmed by difficulties they consider out of control,

- *suicide_risk* shows if a person is considered to be a suicide risk. The value 0 refers to persons who are not considered to be suicide risks while the value 1 refers to persons who are considered to be suicide risks,
- *vicarious* music refers to when somebody listens to music when angry and/or bringing out aggressiveness by listening to music,
- *worshipping* refers to behavioural manifestation of worshiping (e.g. hanging posters, acquiring information about singers, hanging out with other fans),

Note: For all the previous variables that are measured in a scale (i.e. *father_negligence, mother_negligence, isolation, meaninglessness, metal, normlessness, self_estrangement, vicarious, worshipping*); high values of the scale correspond to a behaviour/feeling that happens more, while low values correspond to a behaviour/feeling that is less present.

The structure of the data can be seen below.

```
## Data set with 121 obs. of 15 variables:
##  $ Age              : Itvl. item  num  15.8 14.9 15.3 15.8 14.9 ...
##  $ Age_Group        : Nmnl. item w/ 2 labels for 1,2  num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Drug_Use         : Itvl. item  num  8 9 5 11 7 4 5 7 5 4 ...
##  $ Father_Negligence: Itvl. item + ms.v.  num  17 23 15 11 13 29 10 27 23 12 ...
##  $ Gender           : Nmnl. item w/ 2 labels for 1,2 + ms.v.  num  1 1 1 1 1 1 1 1 1
##  $ Isolation        : Itvl. item  num  6 8 18 9 5 15 8 6 10 5 ...
##  $ Marital_Status   : Nmnl. item w/ 2 labels for 1,2 + ms.v.  num  1 1 1 2 1 2 1 2 1
##  $ Meaninglessness  : Itvl. item  num  10 26 19 13 13 18 12 18 29 22 ...
##  $ Metal            : Itvl. item  num  4.84 6 6 4 8 ...
##  $ Mother_Negligence: Itvl. item  num  10 12 16 10 16 18 9 12 21 15 ...
##  $ Normlessness     : Itvl. item  num  6 8 7 5 3 5 6 7 4 7 ...
##  $ Self_Estrangement: Itvl. item  num  15 20 17 12 6 15 10 12 28 7 ...
##  $ Suicide_Risk     : Nmnl. item w/ 2 labels for 0,1 + ms.v.  num  0 0 0 0 0 1 0 0 0
##  $ Vicarious        : Itvl. item  num  5 4 6 3 3 2 3 3 8 5 ...
##  $ Worshipping      : Itvl. item  num  4 6 3 3 9 4 4 4 9 9 ...
```

Question(s) of interest

The main questions of interest are:

- Is there a relationship between the previous variables and the *self_enstrangement* variable within the data set?
- Which model would you use to fit this data and why? (i.e. what is your main purpose of fitting this model? inference or prediction?)

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.

6.5 Number of extramarital affairs (within the last year)

Data available

This project uses data from Fair (1978). The data refer to a survey that was conducted in the 1960s and 1970s where a questionnaire on sex was published in two different magazines. Readers were asked to mail their answers. The questionnaires included questions about extramarital affairs as well as various demographic and economic characteristics of the individual.

Install the package `AER` and load the data set in R with the command `data(Affairs)`. A data frame containing 601 observations on 9 variables.

- *affairs* numeric. How often engaged in extramarital sexual intercourse during the past year?
- *gender* factor indicating gender.
- *age* numeric variable coding age in years: 17.5 = under 20, 22 = 20–24, 27 = 25–29, 32 = 30–34, 37 = 35–39, 42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 or over.
- *yearsmarried* numeric variable coding number of years married: 0.125 = 3 months or less, 0.417 = 4–6 months, 0.75 = 6 months–1 year, 1.5 = 1–2 years, 4 = 3–5 years, 7 = 6–8 years, 10 = 9–11 years, 15 = 12 or more years.
- *children* factor. Are there children in the marriage?
- *religiousness* numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly, 4 = somewhat, 5 = very.
- *education* numeric variable coding level of education: 9 = grade school, 12 = high school graduate, 14 = some college, 16 = college graduate, 17 = some graduate work, 18 = master's degree, 20 = Ph.D., M.D., or other advanced degree.
- *occupation* numeric variable coding occupation according to Hollingshead classification (reverse numbering).
- *rating* numeric variable coding self rating of marriage: 1 = very unhappy, 2 = somewhat unhappy, 3 = average, 4 = happier than average, 5 = very happy.

The structure of the data can be seen below.

```
## 'data.frame':    601 obs. of  9 variables:
##  $ affairs      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gender       : Factor w/ 2 levels "female","male": 2 1 1 2 2 1 1 2 1 2 ...
##  $ age          : num  37 27 32 57 22 32 22 57 32 22 ...
##  $ yearsmarried : num  10 4 15 15 0.75 1.5 0.75 15 15 1.5 ...
##  $ children     : Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 1 2 2 1 ...
##  $ religiousness: int  3 4 1 5 2 2 2 2 4 4 ...
##  $ education    : num  18 14 12 18 17 17 12 14 16 14 ...
##  $ occupation   : int  7 6 1 6 6 5 1 4 1 4 ...
##  $ rating       : int  4 4 4 5 3 5 3 4 2 5 ...
```

Question(s) of interest
The main questions of interest are:

- Is there a relationship between the previous variables and the number of extramarital affairs a person had?
- Which model would you use to fit this data and why? (i.e. what is your main purpose of fitting this model? inference or prediction?)

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.

6.6 Number of shipping accidents

Data available
The data contains values on the number of reported accidents for ships belonging to a company over a given time period.

The data set, titled `ships`, is located within the `MASS` package. A data frame with 40 observations on the following 7 variables.

- *type* type: "A" to "E".
- *year* year of construction: 1960–64, 65–69, 70–74, 75–79 (coded as "60", "65", "70", "75").
- *period* period of operation : 1960–74, 75–79.
- *service* aggregate months of service.
- *incidents* number of damage incidents.

Question(s) of interest
The main questions of interest are:

- Is there a relationship between the previous variables and the *self_enstrangement* variable within the data set?
- Which model would you use to fit this data and why? (i.e. what is your main purpose of fitting this model? inference or prediction?)

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.

7 How does the weather affect air pollution levels? (1)

Project level: Moderate

7.1 Overall project description

Air quality is an important Public Health issue across the UK with it being a particularly large issue in our biggest city of London. Air pollution is worst in the centre of London where there are many vehicles on the road, these areas also have a large number of pedestrians being exposed to the pollution and so they are areas of primary interest in many studies. However, there are also issues with air pollution in the more suburban regions of London which are surrounded by busy roads like the M25 and industrial sites. These suburban regions will be the focus in this study, specifically the London Borough of Barking & Dagenham which is situated in North London.

Many factors contribute to air pollution like traffic and industry. What the weather is doing also has a direct effect on the concentrations of pollution in the air. Prevailing weather conditions can weaken or improve air quality, for example, strong winds can quickly transport pollutants hundreds of kilometres whereas during calmer conditions, pollutants can accumulate around the source of the release. The climate in the UK is very variable and so in this project you will work with data on air pollution concentrations and weather variables to investigate which specific variables associated with the weather are related to air pollution levels in these suburban regions of London. We will assume for the duration of this project the the amount of pollution being emitted into the air on a daily basis remains relatively constant.

7.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Daily data are available on air pollution and weather variables for 2010. For details see http://www.londonair.org.uk/LondonAir. The data are stored in `LDNdata.csv` and contain the following columns.

- `DATE` - The day number i.e. January 1st = 1 to Deceomber 31st = 365
- `NOX` - Measure of air pollution ($\mu g m^{-3}$)
- `RAIN` - Total rainfall (mm)
- `TEMP` - Mean temperature (^{o}C)
- `SOLR` - Mean solar radiation ($W m^{-2}$)
- `BP` - Mean barometric pressure ($mBar$)
- `WSPD` - Mean wind speed (ms^{-1})

Question(s) of interest

The main questions of interest are:

- Which weather variables affect air pollution levels?
- What is the temporal trend in air pollution?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Flexible regression / Time Series.

8 Investigating free school meals in Scotland (1)

Project level: Moderate

8.1 Overall project description

For some families in Scotland, paying for childrens school meals can be a financial strain. A family with two young children spends approximately ?685 a year on school lunches. To help ease the financial burden on families and increase their disposable income the Scottish Government announced that, as of January 2015, all children in Primary 1, 2 and 3 in Scotland would be entitled to a healthy free school lunch. However, parents who have children in Primary 4 and up need to meet specific personal finance conditions for their children to be eligible for free school meals, the criteria can be found at https://www.mygov.scot/school-meals/.

The universal provision of free, healthy school meals has proven benefits in relation to uptake, family budgets, educational attainment and addressing inequality. In this project you will investigate the association between school level variables and the proportion of pupils who are registered for free school meals. It is well known that poverty and free school meals are closely related however here you will also investigate the effect of other variables associated with a schools location. For this study you will focus on publicly funded primary (primary 4-7) and secondary schools.

8.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Demographic and location data from 2278 primary and secondary schools in Scotland are available, these data were recorded in 2018. This data has been provided by the Scottish Government, see https://www2.gov.scot/Topics/Statistics/Browse/School-Education/Datasets for further details. The data are stored in `schools-project.csv` and contain the following columns.

- `PostCode` - Schools postcode
- `Local.Authority` - Local authority in charge of the school
- `Name` - Schools name
- `Type` - Primary or Secondary
- `free.meals` - Percentage of pupils registered for free school meals, for primary schools this is only for primary 4-7
- `rural.urban` - 6-fold rural/urban measure (see https://www2.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification for further details.)

- `Condition` - Condition of school (A: Good - Performing well and operating efficiently, B: Satisfactory - Performing adequately but showing minor deterioration, C: Poor - Showing major defects and/or not operating adequately, D: Bad - Economic life expired and/or risk of failure)
- `Suitability` - Suitability of school (A: Good - Performing well and operating efficiently, B: Satisfactory - Performing adequately but with minor problems, C: Poor - Showing major problems and/or not operating optimally, D: Bad - Does not support the delivery of services to children and communities)
- `SIMD` - Scottish Index of Multiple Deprivation quintile
- `Easting` - Easting coordinate of school
- `Northing` - Northing coordinate of school
- `No.FTE.teachers` - Number of FTE teachers

Question(s) of interest

The main questions of interest are:

- Which variables are associated with the percentage of school children recieving free school meals?
- Do these predictor variables differ for different subpopulations?
- Are there groups of schools with similar characteristics? Is there a specific variable driving these groupings?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression models.
- Multivariate methods.
- Spatial statistics.
- Flexible regression.

9 Neighbourhood level crime in Baltimore (1)

Project level: Moderate

9.1 Overall project description

The city of Baltimore is infamous for its high crime rates and is often ranked as one of the top 10 most dangerous cities in the US. In 2015 violent crime rates spiked in the city after protests following the death of Freddie Gray in police custody. It is clearly of interest to understand better Baltimore's crime rates and thus, this project will look at understanding how different crime rates vary across the 55 neighbourhoods of Baltimore and how these rates have changed over time.

9.2 Individual project details

How many individual projects are available in this area: 1.

Data available
The Baltimore Neighbourhood Indicator Alliance (BNIA) was set up in 2000 and is "dedicated to producing reliable and actionable quality of life indicators for Baltimore's neighborhood" https://bniajfi.org/. The BNIA annually collect "Vital Signs" which are groups of related data points compiled from a variety of reliable sources that "take the pulse" of Baltimore's neighborhoods. Your focus in this project will be on data from the "Crime and Safety" group in 2011 and 2015. The data are stored in `crime-data.csv` and contains the following columns for each neighbourhood. Note: all rates are reported per 1000 residents.

- `X` - Neighbourhood
- `crime11` - The Part 1 crime rate for 2011
- `crime15` - The Part 1 crime rate for 2015
- `viol11` - The violent crime rate for 2011
- `viol15` - The violent crime rate for 2015
- `juvviol11` - The Juvenile arrest rate for violent crimes for 2011
- `juvviol15` - The Juvenile arrest rate for violent crimes for 2015
- `narc11` - Narcotics calls for service (911 calls) rate for 2011
- `narc15` - Narcotics calls for service (911 calls) rate for 2015

Two further files `crime-shape.shp` and `crime-shape.dbf` are also available and can be used to plot maps of Baltimore's neighbourhoods. To work with these shapefiles you will need the `shapefiles`, `spdep` and `CARBayes` libraries in R.

Question(s) of interest
The main questions of interest:

- are their neighbourhoods with similar crime rates for the 4 variables of interest? (i.e. clusters) Are these similarities the same if each crime rate is treated individually?
- do these groups of neighbourhoods with similar statistics change when you compare 2011 to 2015?
- how are crime rates spatially distributed across the city i.e. what does a map of the crime rates look like? and are there any neighbourhoods whose crime rates have change siginificantly over time?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate methods.
- Spatial statistics.

10 Socioeconomic and demographic factors affecting life expectancy (1)

Project level: Moderate

10.1 Overall project description

Life expectancy at birth is a measure of how long a newborn can expect to live assuming they experience the currently prevailing rates of death throughout their life. There is a large amount of variation in life expectancy across the world and there are many factors which can affect a countrys life expectancy. However, variations in life expectancy are not limited to country-level, large varaitions in life expectancy have also been seen as locally as at a neighbourhood-level within a city.

Baltimore, Maryland is an example of this where the life expectancy in neighbourhoods across the city differ significantly. In 2016 babies born in some areas were expected to live up to almost 20 years longer than in others. Given these large variations across the city, the city government are interested in determining which socioeconomic and demographic factors are driving these differences in life expectancy. In this project you will investigate the effect of many different socioeconomic and demographic factors on the life expectancy in neighbourhoods across Baltimore, with the data from this study being recorded in 2016.

10.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The Baltimore Neighbourhood Indicator Alliance (BNIA) was set up in 2000 and is "dedicated to producing reliable and actionable quality of life indicators for Baltimore's neighborhood" https://bniajfi.org/. The BNIA annually collect "Vital Signs" which are groups of related data points compiled from a variety of reliable sources that "take the pulse" of Baltimore's neighborhoods. Your focus in this project will be on data from several groups 2016. The data are stored in `balt-life-exp.csv` and contains the following columns for each neighbourhood.

- `LifeExp` - Average number of years a newborn can expect to live
- `RDI` - Racial Diversity Index (the percent chance that two people picked at random will be of the same race/ethnicity)
- `AvgHHSize` - Average household size
- `MedIncome` - Median household income
- `PercBelowPovLine` - Percentage of famillies living below the poverty line
- `CrimeRate` - part 1 crime rate per 1000 residents
- `HSDropOut` - Percentage of 9th-12th graders who withdrew from school

- `PercTANF` - Percentage of famillies recieving TANF (Temporary Assistance for Needy Famillies)
- `InfMort` - Number of infant (babies under 1 year old) deaths per 1000 live births
- `UnempRate` - Percentage of people aged 16-64 that are not currently working (but are looking)
- `PerBatchDeg` - Percentage of people aged 25 or over with a Batchelors degree
- `NonLabour` - Percentage of people not in the labor force aged 16-64 (i.e. persons who are not working due to disability, they are in education etc.)
- `PercNoVeh` - Percentage of households with no personal vehicle access

Shapefiles for the neighbourhoods in Baltimore have also been provided (`life-16.shp`). To work with these shapefiles you will need the `shapefiles`, `spdep` and `CARBayes` libraries in R.

Question(s) of interest
The main questions of interest:

- Which socioeconomic and demographic variables affect life expectancy in Baltimore?
- Is there an obvious trend in life expectancy geographically?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression models.
- Flexible regression.
- Spatial statistics.

11 Lake water Quality (1)

Project level: Moderate

11.1 Overall project description

Phosphorus is an essential element for plant life and is also commonly found in fertilizers, manure, and organic waste. However, phosphorus can also be a pollutant, with too much of it causing a rapid increase in the growth of algae on the surface of our fresh waters. This creates a thick blanket over the surface which restricts the penetration of sunlight and limits the waters source of oxygen, this in turn affects the aquatic life below the surface. Phosphorus levels can increase naturally however the increases are generally enhanced by human activity.

In this project you will investigate the effect of different lake and landscape variables on the total phosphorus levels of lakes across 3 lake-rich states in the US.

11.2 Individual project details

How many individual projects are available in this area: 1.

Data available
Data were collected from 3 lake-rich states in the US namely Michigan, Maine and Wisconsin. These data were compiled from databases maintained by state agencies responsible for monitoring lakes under the Federal Clean Water Act. Many water chemistry and catchment variables are available, these are detailed below. Here you will be analysing data recorded in 2004, these data are stored in `lake-quality-04.csv`.

- `State` - US state lake is located in
- `lake` - name of the lake
- `X` - Easting of lake centroid
- `Y` - Northing of lake centroid
- `YEAR` - Year sample was taken
- `area.m2` - surface area calculated using GIS polygons (m^2)
- `per.m` - perimeter (m)
- `elev.m` - elevation (m)
- `mean.depth.m` - mean depth (m)
- `TP.ugL` - Total Phosphorus (ug/L) (measures all forms of phosphorus, both dissolved and particulate.)
- `col.PtCo` - True water colour measured in platinum cobalt units
- `urban` - Percentage of 500m buffer region that was urban
- `forest` - Percentage of 500m buffer region that was forest
- `agri` - Percentage of 500m buffer region that was agricultural (both pasture and crops)

- `wetland` - Percentage of 500m buffer region that was wetland
- `bare` - Percentage of 500m buffer region that was bare (bare rock, sand and clay)

Question(s) of interest
The main questions of interest:

- What relationships are evident between lake variables and the total phosphorus?
- What evidence is available that common patterns exist over space?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression models.
- Flexible regression.
- Spatial statistics.

12 Modelling obesity in Scotland (5)

Project level: Easy/Moderate

12.1 Overall project description

The prevalence of obesity in Scotland has been monitored since the introduction of the Scottish Health Survey, which is designed to monitor the health of the Scottish population living in private households. The main aim of the survey is to keep an eye on health trends in Scotland. The Scottish Heath Survey data will be used to explore trends in obesity in Scotland by examining the Body Mass Index (BMI), which is used to estimate the ideal body weight of an individual. Despite its flaws, the BMI is the most widely used indicator of weight categorisation due to the ease with which it can be attained in large population studies. Differences in obesity prevalence by age, gender, socio-economic status and lifestyle factors will also be investigated.

12.2 Individual project details

How many individual projects are available in this area: 5.

12.3 Obesity prevalence and the BMI distribution (Project 1)

Data available

Data are available on the BMI and socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `ObesityProject1.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in BMI by age, gender, socio-economic status or lifestyle factors?
- Are these differences in the BMI distribution the same across its entire distribution?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear models.

12.4 Obesity prevalence and the BMI distribution (Project 2)

Data available
Data are available on the BMI and socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `ObesityProject2.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest
The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in BMI by age, gender, socio-economic status or lifestyle factors?
- Are these differences in the BMI distribution the same across its entire distribution?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear models.

12.5 Obesity prevalence and the BMI distribution (Project 3)

Data available
Data are available on the BMI and socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `ObesityProject3.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)

- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest
The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in BMI by age, gender, socio-economic status or lifestyle factors?
- Are these differences in the BMI distribution the same across its entire distribution?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear models.

12.6 Obesity prevalence and weight categorisation (Project 1)

Data available
Data are available on the BMI and socio-economic and lifestyle factors from the 2008 - 2011 Scottish Health Surveys. The data are stored in `ObesityProject4.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual
- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest
The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in BMI by age, gender, socio-economic status or lifestyle factors?
- Are there any differences in the BMI weight classification groups by age, gender, socio-economic status or lifestyle factors?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear models.

12.7 Obesity prevalence and weight categorisation (Project 2)

Data available

Data are available on the BMI and socio-economic and lifestyle factors from the 2012 - 2016 Scottish Health Surveys. The data are stored in `ObesityProject5.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual
- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in BMI by age, gender, socio-economic status or lifestyle factors?
- Are there any differences in the BMI weight classification groups by age, gender, socio-economic status or lifestyle factors?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear models.

13 Modelling the progression of world records in athletics (3)

Project level: Easy/Moderate

13.1 Overall project description

The International Association of Athletics Federations (IAAF) is the international governing body for athletics. It was founded on 17 July 1912 as the International Amateur Athletic Federation. Since that date, the IAAF has been the body that ratifies (or not) claims to world records in all athletic disciplines. These three projects will use official datasets from the IAAF in an attempt to model the way world records in various events have progressed over time. They will also compare the rates of progress in some different (but comparable) events.

13.2 Individual project details

How many individual projects are available in this area: 3.

13.3 Events over short distances

Data available

Data are available on the progression of world record times for the following events - 100 metres sprint, 200 metres sprint and 110 metres hurdles for men and 100 metres sprint, 200 metres sprint and 100 metres hurdles for women. The data are stored in `Men100m.csv`, `Men200m.csv`, `Men110hurdles.csv`, `Women100m.csv`, `Women200m.csv` and `Women100hurdles.csv`. Each file contains the following eight pieces of information for every ratified occasion on which the previous world record was beaten:

- `Index` - A serial number from 1 to n.
- `Time` - The new world record time (in seconds).
- `Wind` - Wind speed (m/s) measured parallel to the athlete's direction of travel; a positive value means the wind was in the same direction as the athlete's run while a negative value means the wind was against the athlete.
- `Competitor` - The name of the new world record holder.
- `DOB` - The new world record holder's date of birth (dd/mm/yyyy).
- `Country` - The country that the new world record holder represented (a 3-letter code).
- `Venue` - Where the new world record was set.
- `Date` - The date when the new world record was set (dd/mm/yyyy).

Question(s) of interest

The main questions of interest are:

- For each event separately, fit and assess a linear trend to the world record times. [It is not reasonable to expect a linear model to fit these data since, by extrapolation, that would suggest athletic performance could continue to improve without limit.]
- If the linear model is not appropriate, try fitting linear models to various transformations of the record time - for example, the log world record time, the percentage of new to very first recorded world record time, new world record speed for this event.
- Does one of the linear models fit better if the world record is adjusted for wind speed?
- Fit and assess a generalised additive model (GAM) to the trend in world record times.
- Use the best-fitting model type to compare the patterns of progress in the three events for men and women separately.
- Use the best-fitting model type to compare the patterns of progress for men and women in each event separately.

Relevant courses
In undertaking this project, you might find it helpful to have taken the following course:

- Flexible regression.

13.4 Events over middle distances

Data available
Data are available on the progression of world record times for the following events - 400 metres, 800 metres and 1500 metres for both men and women. The data are stored in `Men400m.csv`, `Men800m.csv`, `Men1500m.csv`, `Women400m.csv`, `Women800m.csv` and `Women1500m.csv`. Each file contains the following eight pieces of information for every ratified occasion on which the previous world record was beaten:

- `Index` - A serial number from 1 to n.
- `Time` - The new world record time (in seconds).
- `Wind` - Wind speed (m/s) measured parallel to the athlete's direction of travel; a positive value means the wind was in the same direction as the athlete's run while a negative value means the wind was against the athlete. [This is rarely recorded for distances above 200m.]
- `Competitor` - The name of the new world record holder.
- `DOB` - The new world record holder's date of birth (dd/mm/yyyy).
- `Country` - The country that the new world record holder represented (a 3-letter code).
- `Venue` - Where the new world record was set.
- `Date` - The date when the new world record was set (dd/mm/yyyy).

Question(s) of interest
The main questions of interest are:

- For each event separately, fit and assess a linear trend to the world record times. [It is not reasonable to expect a linear model to fit these data since, by extrapolation, that would suggest athletic performance could continue to improve without limit.]
- If the linear model is not appropriate, try fitting linear models to various transformations of the record time - for example, the log world record time, the percentage of

new to very first recorded world record time, new world record speed for this event.

- Fit and assess a generalised additive model (GAM) to the trend in world record times.
- Use the best-fitting model type to compare the patterns of progress in the three events for men and women separately.
- Use the best-fitting model type to compare the patterns of progress for men and women in each event separately.

Relevant courses

In undertaking this project, you might find it helpful to have taken the following course:

- Flexible regression.

13.5 Events over longer distances

Data available

Data are available on the progression of world record times for the following events - 1500 metres, 5000 metres and 10000 metres for both men and women. The data are stored in `Men1500m.csv`, `Men5000m.csv`, `Men10000m.csv`, `Women1500m.csv`, `Women5000m.csv` and `Women10000m.csv`. Each file contains the following eight pieces of information for every ratified occasion on which the previous world record was beaten:

- `Index` - A serial number from 1 to n.
- `Time` - The new world record time (in seconds).
- `Wind` - Wind speed (m/s) measured parallel to the athlete's direction of travel; a positive value means the wind was in the same direction as the athlete's run while a negative value means the wind was against the athlete. [This is rarely recorded for distances above 200m.]
- `Competitor` - The name of the new world record holder.
- `DOB` - The new world record holder's date of birth (dd/mm/yyyy).
- `Country` - The country that the new world record holder represented (a 3-letter code).
- `Venue` - Where the new world record was set.
- `Date` - The date when the new world record was set (dd/mm/yyyy).

Question(s) of interest

The main questions of interest are:

- For each event separately, fit and assess a linear trend to the world record times. [It is not reasonable to expect a linear model to fit these data since, by extrapolation, that would suggest athletic performance could continue to improve without limit.]
- If the linear model is not appropriate, try fitting linear models to various transformations of the record time - for example, the log world record time, the percentage of new to very first recorded world record time, new world record speed for this event.
- Fit and assess a generalised additive model (GAM) to the trend in world record times.
- Use the best-fitting model type to compare the patterns of progress in the three events for men and women separately.
- Use the best-fitting model type to compare the patterns of progress for men and women in each event separately.

Relevant courses

In undertaking this project, you might find it helpful to have taken the following course:

- Flexible regression.

14 Modelling the growth of infants in Glasgow (3)

Project level: Moderate

14.1 Overall project description

Some years ago, a researcher recruited a cohort of babies born in a Glasgow maternity hospital for a growth study. She recruited 127 babies and measured them at birth (0 months), 1, 2, 3, 4, 5, 6, 9, 12, 18 and 24 months. Particularly important measurements are the infants' lengths (or heights), head circumferences and weights. Body mass index, which crudely adjusts weight for length, is obtained by BMI = weight/length^2. An unusual feature of the dataset is that one researcher made all the measurements, which eliminates the inter-rater variability in measuring length and weight which is an important component of measurement error in most research studies of this kind. The researcher also recorded a number of pieces of information about the feeding of the infant in early life, the infant's family (such as an index of social deprivation) and the infant's mother (such as smoking behaviour during the pregnancy). The researcher was able to follow up almost all of the infants at almost all the time points, so there are few missing values in the dataset. These projects will use the data to model the growth of Glasgow children during the first two years of their life and explore the extent to which background variables affect patterns of growth.

14.2 Individual project details

How many individual projects are available in this area: 3.

14.3 Growth in length during infancy

Data available

Data are available on the length of each infant at each of the time points; these are stored in `Length.csv`. The file contains the following 14 pieces of information for every length measurement:

- `Subject` - A serial number from 1 to 127.
- `Age` - The infant's age (in months).
- `Length` - The infant's length or height (in cms).
- `Gender` - The infant's gender: 0 = boy; 1 = girl.
- `Feed Type` - How the infant was fed as a baby: 0 = exclusively breast-fed; 1 = at least partially bottle fed.
- `Duration BF` - For how long the infant was (at least partially) breast-fed (months).
- `Age Solids` - The age of the infant when introduced to solid food (months).
- `Dep Cat` - A social deprivation score for the infant's family: a number from 1 to 7, where 1 is the least deprived category.
- `M Age` - The mother's age at the infant's birth (years).

- `M FHE` - The mother's level of education: 0 = not FE/HE; 1 = FE/HE.
- `M Prev` - Whether or not the mother had a previous child or children: 0 = No; 1 = Yes.
- `M Smoke` - Whether or not the mother smoked during pregnancy: 0 = No; 1 = Yes.
- `M Height` - Mother's height (cms)
- `F Height` - Father's height (cms)

Question(s) of interest

The main questions of interest are:

- Fit, and assess the fit of, a simple linear model of length on age in infancy.
- If the linear model is not appropriate, try fitting linear models to various transformations of the length - for example, the logarithm or square root. Try a low-order polynomial, for example a quadratic in age.
- Use the other explanatory variables to try to improve the fit of your model.
- Fit, and assess the fit of, a generalised additive model (GAM) to the relationship between length and age.
- Use the other explanatory variables to try to improve the fit of your GAM.
- Fit mixed models, either linear mixed models or generalised additive mixed models, with a random effect of subject.

Relevant courses

In undertaking this project, you might find it helpful to have taken the following courses:

- Flexible regression.
- Linear mixed models.

14.4 Growth in head circumference during infancy

Data available

Data are available on the head circumference of each infant at each of the time points; these are stored in `HeadCircumference.csv`. The file contains the following 14 pieces of information for every length measurement:

- `Subject` - A serial number from 1 to 127.
- `Age` - The infant's age (in months).
- `HC` - The infant's head circumference (in cms).
- `Gender` - The infant's gender: 0 = boy; 1 = girl.
- `Feed Type` - How the infant was fed as a baby: 0 = exclusively breast-fed; 1 = at least partially bottle fed.
- `Duration BF` - For how long the infant was (at least partially) breast-fed (months).
- `Age Solids` - The age of the infant when introduced to solid food (months).
- `Dep Cat` - A social deprivation score for the infant's family: a number from 1 to 7, where 1 is the least deprived category.
- `M Age` - The mother's age at the infant's birth (years).
- `M FHE` - The mother's level of education: 0 = not FE/HE; 1 = FE/HE.

- `M Prev` - Whether or not the mother had a previous child or children: 0 = No; 1 = Yes.
- `M Smoke` - Whether or not the mother smoked during pregnancy: 0 = No; 1 = Yes.
- `M Height` - Mother's height (cms)
- `F Height` - Father's height (cms)

Question(s) of interest
The main questions of interest are:

- Fit, and assess the fit of, a simple linear model of head circumference on age in infancy.
- If the linear model is not appropriate, try fitting linear models to various transformations of the head circumference - for example, the logarithm or square root. Try a low-order polynomial, for example a quadratic in age.
- Use the other explanatory variables to try to improve the fit of your model.
- Fit, and assess the fit of, a generalised additive model (GAM) to the relationship between head circumference and age.
- Use the other explanatory variables to try to improve the fit of your GAM.
- Fit mixed models, either linear mixed models or generalised additive mixed models, with a random effect of subject.

Relevant courses
In undertaking this project, you might find it helpful to have taken the following courses:

- Flexible regression.
- Linear mixed models.

14.5 Development of body mass index during infancy

Data available
Data are available on the body mass index (BMI) of each infant at each of the time points; these are stored in `BMI.csv`. The file contains the following 14 pieces of information for every length measurement:

- `Subject` - A serial number from 1 to 127.
- `Age` - The infant's age (in months).
- `BMI` - The infant's body mass index (in kg/m^2).
- `Gender` - The infant's gender: 0 = boy; 1 = girl.
- `Feed Type` - How the infant was fed as a baby: 0 = exclusively breast-fed; 1 = at least partially bottle fed.
- `Duration BF` - For how long the infant was (at least partially) breast-fed (months).
- `Age Solids` - The age of the infant when introduced to solid food (months).
- `Dep Cat` - A social deprivation score for the infant's family: a number from 1 to 7, where 1 is the least deprived category.
- `M Age` - The mother's age at the infant's birth (years).
- `M FHE` - The mother's level of education: 0 = not FE/HE; 1 = FE/HE.
- `M Prev` - Whether or not the mother had a previous child or children: 0 = No; 1 = Yes.

- `M Smoke` - Whether or not the mother smoked during pregnancy: 0 = No; 1 = Yes.
- `M Height` - Mother's height (cms)
- `F Height` - Father's height (cms)

Question(s) of interest
The main questions of interest are:

- Fit, and assess the fit of, a simple linear model of BMI on age in infancy.
- If the linear model is not appropriate, try fitting linear models to various transformations of the BMI - for example, the logarithm or square root. Try a low-order polynomial, for example a quadratic in age.
- Use the other explanatory variables to try to improve the fit of your model.
- Fit, and assess the fit of, a generalised additive model (GAM) to the relationship between BMI and age.
- Use the other explanatory variables to try to improve the fit of your GAM.
- Fit mixed models, either linear mixed models or generalised additive mixed models, with a random effect of subject.

Relevant courses
In undertaking this project, you might find it helpful to have taken the following courses:

- Flexible regression.
- Linear mixed models.

15 Identifying Seasonal Coherence in Global Lake Surface Water Temperature (1)

Project level: Moderate

15.1 Overall project description

The quantity of data we are collecting is increasing at an unprecedented rate with the advent of new Earth Observation (EO) technologies that obtain data on our environment using satellites. These new data sets enable us to use statistical models to explore and describe changes in our natural environment.

It is often of interest to explore the coherence in environmental variables via a clustering method which can be used to identify groups of individuals which share similar characteristics. By identifying which groups of individuals that are behaving in a similar way we can then look for drivers of common patterns. Specifically, this project will investigate temporal coherence, which is defined as the synchrony between major fluctuations in a set of time series, in the surface water temperature at a set of lakes across the globe.

15.2 Individual project details

How many individual projects are available in this area: 1.

Data available
For 30 large lakes across the globe there is a time series of bi-monthly average lake surface water temperature (LSWT). All measurements have been obtained from the ARCLake project (www.laketemp.net) which uses information from the European Space Agency's AATSR instrument on board the MERIS satellite platform. Another data has been available by the Globolakes project (www.globolakes.ac.uk). The data available cover the time period from January 2003 until December 2011. There are two data sets available.

The first `arcdata.csv` contains time series of temperature with columns corresponding to different lakes.

- `date` - The date in decimal year format.
- `month` - The month of year.
- `Lake1 ... Lake30` - The lake surface water temperature (LSWT) for Lake 1 to Lake 30 (in Celsius)

The second, named `arcinfo.csv` contains additional information on the lakes in the following columns

- `lakeid` - ID number for the lake (in format Lake1... Lake30)
- `lakename` - name of the lake

- `lat` - latitude
- `long` - longitude
- `elevation` - elevation of the lake (m)

Question(s) of interest
The main questions of interest are:

- To estimate the average seasonal pattern at each of the lakes.
- To identify coherent groups of lakes in terms of their seasonal patterns of LSWT and investigate the statistically optimal number of groups required.
- Using the location and elevation data provided to informally explore any drivers of the differences in these groups.

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.
- Linear Models.
- Functional Data Analysis.
- Flexible Regression.

16 Temporal Patterns in Temperature and NDVI (2)

Project level: Moderate

16.1 Overall project description

The quantity of data we are collecting is increasing at an unprecedented rate with the advent of new Earth Observation (EO) technologies that obtain data on our environment using satellites. These new data sets enable us to use statistical models to explore and describe changes in our natural environment. Natural environment variables of interest include temperature on land and water, cloud cover, soil moisture and satellite derived vegetation indices such as the Normalized Difference Vegetation Index (NDVI). NDVI is a indicator of the quantity of live green vegetation in an area and ranges from -1.0 (corresponding to a cloudy or snow covered area) to +1.0 (corresponding to a dense green canopy).

Temporal changes in environmental variables are often complex and we need to account for both long term trends and seasonal patterns within our models. This project will focus on exploring the presence and strength of changes in environmental variables over time and their interactions with other variables.

16.2 Individual project details

How many individual projects are available in this area: 2.

16.3 Toorale National Park

Data available

A monthly time series of land surface temperature (LST) and normalized difference vegetation index (NDVI) are available for an area near Toorale National Park in Australia, a location which is thought to be sensitive to changes in climate.

All measurements have been obtained via the AATSR instrument on board the European Space Agencies MERIS platform and cover the time period from August 2002 until March 2012. The data are stored in a csv file called `australia.csv` which has the following columns;

- `month` - The month of obersvation (1-12).
- `NDVI` - NDVI.
- `LST` - Land Surface Temperature (in Celsius).

Question(s) of interest

The main questions of interest are;

- What are the temporal patterns (trend and season) in NDVI?
- Is there any relationship between NDVI and LST?

- Given the data available what model is optimal in terms of estimating future predictions of NDVI? Use this model to estimate predictions of NDVI up to two years after the end of the time period covered by the data.

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Time Series.
- Environmental statistics.
- Flexible regression.

16.4 Lake Balaton Catchment

Data available
A monthly time series of land surface temperature (LST), lake surface water temperature (LSWT), rainfall and NDVI are available for Lake Balton, Hungary and it's surrounding catchement.

All temperature and NDVI measurements have been obtained via the AATSR instrument on board the European Space Agencies MERIS platform while rainfall data are part of the Climate Prediction Center (CPC) Unified Precipitation Project that is underway at US National Oceanic and Atmospheric Association (NOAA). The data cover the time period from August 2002 until January 2012. The data are stored in a csv file called `balaton.csv` which has the following columns;

- `month` - The month of obersvation (1-12).
- `year` - The year of observation.
- `ndvi` - NDVI in catchment.
- `lst` - Land Surface Temperature of catchment, LST (in Celsius).
- `rain` - Precipitation in catchment (l/mm^2).
- `lswt` - Lake Surface Water Temperature, LSWT (in Celsius).

Question(s) of interest
The main questions of interest are;

- What are the main temporal patterns (trend and season) in NDVI?
- Is there any relationship between Land Surface Temperature for the catchment and Lake Surface Water Temperature?
- Using the data provided, what is the best model for predicting NDVI in the Lake Balaton Catchment? You should describe your approach to obtaining statistically optimal model here.

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Time Series.
- Environmental statistics.
- Flexible regression.

17 Descriptive and predictive modelling of environmental lake quality and process data (6)

Project level: Moderate

17.1 Overall project description

In order to understand the environment around us, monitor the risks and protect human and animal health, it is essential to have a good and thorough understanding of patterns emerging over time within the environment and of the relationships between key drivers of environmental quality and the associated environmental responses. As part of this process, the European Union has specific directives which outline the quality targets and thresholds that have to be met by member states, and environmental regulators such as the Scottish Environment Protection Agency and the Environment Agency are currently responsible for reporting on environmental quality measures to Europe.

For surface water quality, in particular, the EC water framework directive, and associated directives such as the nitrates directive outline targets and thresholds for nutrient and phytoplankton biomass levels within lakes and rivers in an attempt to protect human, animal and ecosystem health.

The projects below are all related to this general area for lakes within the UK. The data arise as time series data (at different temporal resolutions) and are either from one location or multiple locations within, or throughout, a lake. These projects investigate how to interrogate and analyse such data, across different temporal scales, to identify the temporal patterns and relationships for responses such as total phosphorus, chlorophyll or secchi depth and drivers of water quality such as soluble reactive phosphorus, temperature, silica, nitrate, conductivity and pH. Additionally, these lake processes are influenced by the water temperature, which varies with lake depth, and it is of interest to investigate this process and the drivers such as air temperature, solar irradiance and windspeed using high frequency data.

17.2 Individual project details

How many individual projects are available in this area: 6.

17.3 Temporal patterns and drivers of water quality at Loch Leven

Data available

Data are available for Loch Leven, Kinross, Scotland, on water quality responses, nutrients and temperature from 1988-2007 at the monitoring location Reed Bower within Loch Leven. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Dudley, B. J.; May, L.; Spears, B. M.; Kirika, A. (2013). Loch Leven long-term monitoring data: phosphorus, silica and chlorophyll concentrations and temperature, 1985-2007. NERC Environmental Information Data Centre. https://doi.org/10.5285/2969776d-0b59-4435-a746-da50b8fd62a3

The data are stored in the files `chlaRB5.csv`, `SRPRB5.csv`, `TempRB5.csv`, `condRB5.csv`, `SRSRB5.csv` for chlorophyll$_a$ (chla, a proxy measure of water quality), soluble reactive phosphorus (SRP, a nutrient), water temperature, conductivity and soluble reactive silica (a nutrient), and each file contains the following columns:

- `SAMPLEDATE` - the date of the measurement
- `SITEID` - the site ID, which here is RB5 in all for Reed Bower
- `VALUE` - the measured value of the determinand being recorded
- `DETERMINANDNAME` - the name of the determinand being recorded
- `DETERMINANDUNITS` - the units of measurement for the determinand.

While not necessary, there are further data available on Loch Leven from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/eidc/documents#term=Loch+Leven&page=1

Question(s) of interest
The main questions of interest are:

- What are the temporal patterns for chla, SRP, temperature, conductivity and silica?
- What appears to be the effect of nutrients, such as SRP and silica, and temperature and conductivity on the water quality (measured by proxy as chlorophylla)?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.

17.4 Spatio-temporal water quality patterns at Loch Leven

Data available
Data are available for Loch Leven, Kinross, Scotland, on water quality responses from 1985-2007 at 3 locations across Loch Leven. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Dudley, B. J.; May, L.; Spears, B. M.; Kirika, A. (2013). Loch Leven long-term monitoring data: phosphorus, chlorophyll concentrations, water clarity, 1985-2007. NERC Environmental Information Data Centre. https://doi.org/10.5285/2969776d-0b59-4435-a746-da50b8fd62a3

The data are stored in `TPRB5.csv`, `TPS18.csv`, `TPSD6.csv`, `chlaRB5.csv`, `chlaS18.csv`, `chlaSD6.csv` and `sdepthRB5`, `sdepthS18.csv`, `sdepthSD6.csv` where TP is total phosphorus (a proxy measure of water quality), chla is chlorophylla (a proxy measure of water quality) and sdepth is secchi depth (a measure of water clarity) and each contain the following columns,

- `SAMPLEDATE` - the date of the measurement
- `SITEID` - the site ID, Reed Bower (RB5), South Deeps (SD6), Sluices (Sl8)
- `VALUE` - the measured value of the determinand being recorded
- `DETERMINANDNAME` - the name of the determinand being recorded
- `DETERMINANDUNITS` - the units of measurement for the determinand.

While not necessary, there are further data available on Loch Leven from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/eidc/documents#term=Loch+Leven&page=1

Question(s) of interest
The main questions of interest are:

- What are the temporal patterns for chla, TP and water clarity?
- How do these patterns differ by site?
- How is water quality at LL changing over time?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.
- Linear mixed models.

17.5 Temporal patterns and drivers of water temperature at Bassenthwaite Lake

Data available
Data are available hourly for lake temperature, air temperature, solar irradiance and wind speed data from an automatic water monitoring buoy on Bassenthwaite Lake, a lake in the north west of England for 2008-2011. Measurements were taken every 4 minutes and calculated as hourly averages. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Jones, I.; Feuchtmayr, H. (2017). Data from automatic water monitoring buoy from Bassenthwaite Lake, 2008 to 2011. NERC Environmental Information Data Centre. https://doi.org/10.5285/ce702019-77fe-4ca7-b1d4-a7e4eb6e40c0

The data are stored in `BASS1.csv` which contains the following columns,

- `DateGMT` - the date and time of the measurement
- `Water temperature at 1m` - the water temperature at the surface in degrees celcius
- `Air temperature` - the air temperature at various depths in degrees celcius
- `Pyranometer` - the solar irradiance

- `Wind Speed` - the wind speed

While not necessary, there are further data available for 2012-2015 for this lake and on automatic monitoring buoys for other lake district lakes from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/documents/bf30d6aa-345a-4771-8417-ffbcf8c08c28

Question(s) of interest
The main questions of interest are:

- What is the temporal pattern for water temperature at the surface (i.e. 1m depth)?
- What is the effect of air temperature, solar irradiance, and wind speed on the water temperature at the surface?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.

17.6 Temporal patterns and temperature depth profiles at Bassenthwaite Lake

Data available
Data are available hourly for lake temperature at multiple depths and on air temperature from an automatic water monitoring buoy on Bassenthwaite Lake, a lake in the north west of England for 2008-2011. The lake temperatures are measured in various depths of the lake. Measurements were taken every 4 minutes and calculated as hourly averages. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Jones, I.; Feuchtmayr, H. (2017). Data from automatic water monitoring buoy from Bassenthwaite Lake, 2008 to 2011. NERC Environmental Information Data Centre. https://doi.org/10.5285/ce702019-77fe-4ca7-b1d4-a7e4eb6e40c0

The data are stored in `BASS2.csv` which contains the following columns,

- `DateGMT` - the date and time of the measurement
- `Water temperature at 1m, 2m, 3m, 4m, 5m, 6m, 8m, 10m, 12m, 14m, 16m, 18m` - the water temperature at various depths in degrees celcius
- `Air temperature` - the air temperature at various depths in degrees celcius

While not necessary, there are further data available for 2012-2015 for this lake and on automatic monitoring buoys for other lake district lakes from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/documents/bf30d6aa-345a-4771-8417-ffbcf8c08c28

Question(s) of interest
The main questions of interest are:

- What is the temporal pattern for water temperature at the surface (i.e. 1m depth)?
- How does the water temperature temporal pattern differ by depth?
- What is the effect of air temperature on the water temperature at the surface and how does this differ by depth?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.
- Linear mixed models.

17.7 Temporal patterns and drivers of water quality at Basetnhwaite Lake

Data available
This is a long-term monitoring dataset of surface temperature, surface oxygen, water chemistry and phytoplankton chlorophyll-a from fortnightly sampling by the Centre for Ecology & Hydrology (and previously the Institute of Freshwater Ecology) at Bassenthwaite Lake in Cumbria, England. The data available comprise surface temperature (TEMP) in degree Celsius, surface oxygen saturation (OXYG) in % air-saturation, alkalinity (ALK) in ?g per litre as CaCO3 and pH. Soluble reactive phosphate (PO4P), dissolved reactive silicon expressed as SiO2 (SIO2) and phytoplankton chlorophyll a (TOCA) are all given in ?g per litre. Water samples are based on a sample integrated from 0 to 5 m. Measurements are made from a boat at a marked location (buoy) at the deepest part of the lake. When it was not possible to visit the buoy, samples were taken from the shore, thus water samples were not integrated on these occasions, marked as Flag 2. All data are from August 1990 until the end of 2013. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Maberly, S.C.; Carter, H.T.; Clarke, M.A.; De Ville, M.M.; Fletcher, J.M.; James, J.B.; Keenan, P.; Kelly, J.L.; Mackay, E.B.; Parker, J.E.; Patel, M.; Pereira, M.G.; Rhodes, G. ; Tanna, B.; Thackeray, S.J.; Vincent, C.; Feuchtmayr, H. (2017). Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Bassenthwaite Lake, 1990 to 2013. NERC Environmental Information Data Centre. https://doi.org/10.5285/91d763f2-978d-4891-b3c6-f41d29b45d55

The data are stored in `AlkBass.csv`, `OxyBass.csv`, `PO4PBass.csv`, `SIO2Bass.csv`, `TOCABass.csv`, `TEMPBass.csv`, which each contain the following columns,

- `sdate` - the date of the measurement
- `variable` - the variable being measured
- `value` - the measured value of the determinand being recorded

- `sign_if_LT_LOD` - a sign to indicate values are below the limit of detection
- `flag` - a flag to indicate the location where the sample was recorded

While not necessary, there are further data available on Loch Leven from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/documents/bf30d6aa-345a-4771-8417-ffbcf8c08c28

Question(s) of interest
The main questions of interest are:

- What are the temporal patterns for chlorophyll, alkalinity, phosphorus, oxygen, temperature and silica?
- What appears to be the effect of nutrients, such as phosphorus, silica and alkalinity, and temperature and oxygen on the water quality (measured by proxy as chlorophylla)?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.

17.8 Temporal patterns and drivers of water quality at Derwent Water

Data available
This is a long-term monitoring dataset of surface temperature, water chemistry and phytoplankton chlorophyll a from fortnightly sampling by the Centre for Ecology & Hydrology (and previously the Institute of Freshwater Ecology) at Derwent Water in Cumbria, England. The data available comprise surface temperature (TEMP) in degree Celsius, alkalinity (ALK) in ?g per litre as CaCO3 and pH. Nitrate (NO3N), soluble reactive phosphate (PO4P) and phytoplankton chlorophyll a (TOCA) are all given in ?g per litre. Measurements are made from a boat at a marked location (buoy) at the deepest part of the lake. When it was not possible to visit the buoy, samples were taken from the shore, thus water samples were not integrated on these occasions, marked as Flag 2. All data are from August 1990 until the end of 2013. The original data have been supplied by the Natural Environment Research Council through the Environmental Information Data Centre platform, with full reference: Maberly, S.C.; Carter, H.T.; Clarke, M.A.; De Ville, M.M.; Fletcher, J.M.; James, J.B.; Keenan, P.; Kelly, J.L.; Mackay, E.B.; Parker, J.E.; Patel, M.; Pereira, M.G.; Rhodes, G. ; Tanna, B.; Thackeray, S.J.; Vincent, C.; Feuchtmayr, H. (2017). Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Bassenthwaite Lake, 1990 to 2013. NERC Environmental Information Data Centre. https://catalogue.ceh.ac.uk/documents/106844ff-7b4c-45c3-8b4c-7cfb4a4b953b

The data are stored in `AlkDerw.csv`, `NitrateDerw.csv`, `PHDerw.csv`, `PO4PDerw.csv`, `TOCADerw.csv`, `TEMPDerw.csv`, which each contain the following columns,

- `sdate` - the date of the measurement
- `variable` - the variable being measured
- `value` - the measured value of the determinand being recorded
- `sign_if_LT_LOD` - a sign to indicate values are below the limit of detection
- `flag` - a flag to indicate the location where the sample was recorded

Question(s) of interest

The main questions of interest are:

- What are the temporal patterns for chlorophyll, nitrate, phosphorus, temperature, pH and alkalinity?
- What appears to be the effect of nutrients, such as phosphorus and nitrate, and temperature and alkalinity on the water quality (measured by proxy as chlorophyll)?

While not necessary, there are further data available on Loch Leven from the Environmental Information Data Centre if this is of interest to explore: https://catalogue.ceh.ac.uk/documents/bf30d6aa-345a-4771-8417-ffbcf8c08c28

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Linear models.
- Time series.
- Flexible regression.
- Environmental statistics.

18 Identifying house price clusters in Glasgow (1)

Project level: Moderate

18.1 Overall project description

Cost of housing is one of the largest proportions of individual household spending and the cost or value of dwellings can vary hugely over time. In this context we are interested in the spatial pattern of housing areas in Glasgow. This project will look at cluster analysis of yearly median house price time series data for intermediate zones in the Glasgow City Local Authority region. This can then be used to visualise how the intermediate zones form submarkets.

18.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on median yearly house prices from 1993 to 2013 for each intermediate zone (IZ) in the Glasgow City Local Authority region, which are small spatial areas created for the distribution of small-area statistics. For details see https://statistics.gov.scot/home, and the average population of each IZ is around 4,000 people. The data are stored in `HousePrices.csv` and contain the following columns.

- `Feature Identifier` - A unique code for each IZ area.
- `Feature Name` - The name of the IZ area.
- `1993` - The median house price in each IZ in 1993.
- …
- `2013` - The median house price in each IZ in 2013.

Question(s) of interest

The main questions of interest are:

- Are there clusters of IZs in each year that exhibit similar house prices?
- How do the estimated cluster structures change over time?
- Can you estimate a single cluster structure for all years simultaneously?
- Can you model the time series of house prices in each IZ as smooth curves and cluster these?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.
- Functional Data Analysis.
- Spatial statistics.

19 Cluster Analysis of Acute Respiratory Distress Syndrome (2)

Project level: Moderate

19.1 Overall project description

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 whether groups exist in the biomedical markers data both before and after treatment and whether these clusters connect to the patient's outcome and whether ECMO changes these.

19.2 Individual project details

How many individual projects are available in this area: 2.

Data available
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).

- `Pt_ID` - A unique code for each patient.
- `Gender` - Patient Gender (m=Male, f=Female)
- `Indication` - A disease indicator with the following levels
  - ALF = acute lung failure
  - 1 = viral pneumonia
  - 2 = bacterial pneumonia
  - 3 = aspiration pneumonitis
  - 4 = ARDS Trauma
  - 5 = ARDS surgery
  - 6 = Chemo
  - 7 = other
- ECMO_Survival - a survival indicator, Y= survivor, N = non-survivor (**Do not use this variable for your cluster analysis**, use it to check the cluster analysis results)
- Hospital_Survival - a secondary survival indicator, Y= survivor, N = non-survivor (**Do not use this variable for your cluster analysis**, use it to check the cluster analysis results)
- Duration_ECMO - Days of ECMO treatment
- The following variables all have two variants: PreECMO and Day1ECMO

- RR - Respiratory rate
- Vt - Tidal volume
- FiO2 - Inspired fraction of oxygen
- Ppeak - Peak airway pressure
- Pmean - Mean airway pressure
- PEEP - Positive end expiratory pressure
- PF - Arterial partial pressure of oxygen/inspired fraction of oxygen ratio
- SpO2 - Periperal oxygen saturation
- PaCO2 - Arterial partial pressure of carbon dioxide
- pH - Arterial pH
- BE - Arterial base excess
- Lactate - Arterial lactate
- NAdose - Noradrenaline dose
- MAP - Mean arterial pressure
- Creatinine -
- Urea -
- CK - Creatinine Kinase
- Bilirubin -
- Albumin -
- CRP - C reactive protein
- Fibrinogen -
- Ddimer -
- ATIII - Anti-thrombin III
- HB - Haemaglobin
- Leukocytes -
- Platelets -
- TNFa -
- IL6 -
- IL8 -
- siL2

19.3 PreECMO data

Question(s) of interest
The main questions of interest are:

- Can we find clusters in the PreECMO biomedical markers data?
- Do the clusters found correspond at all to the outcome variables for survival (Hospital_Survival and ECMO_Survival)

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.

19.4 Day1ECMO data

Question(s) of interest

The main questions of interest are:

- Can we find clusters in the Day1ECMO biomedical markers data?
- Do the clusters found correspond at all to the outcome variables for survival (Hospital_Survival and ECMO_Survival)

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.

20 Classification Analysis of Acute Respiratory Distress Syndrome (2)

Project level: Moderate

20.1 Overall project description

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.

20.2 Individual project details

How many individual projects are available in this area: 2.

Data available
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).

- `Pt_ID` - A unique code for each patient.
- `Gender` - Patient Gender (m=Male, f=Female)
- `Indication` - A disease indicator with the following levels
  - ALF = acute lung failure
  - 1 = viral pneumonia
  - 2 = bacterial pneumonia
  - 3 = aspiration pneumonitis
  - 4 = ARDS Trauma
  - 5 = ARDS surgery
  - 6 = Chemo
  - 7 = other
- ECMO_Survival - a survival indicator, Y= survivor, N = non-survivor
- Hospital_Survival - a secondary survival indicator (ignored for this analysis), Y= survivor, N = non-survivor
- Duration_ECMO - Days of ECMO treatment
- The following variables all have two variants: PreECMO and Day1ECMO
  - RR - Respiratory rate
  - Vt - Tidal volume

- FiO2 - Inspired fraction of oxygen
- Ppeak - Peak airway pressure
- Pmean - Mean airway pressure
- PEEP - Positive end expiratory pressure
- PF - Arterial partial pressure of oxygen/inspired fraction of oxygen ratio
- SpO2 - Periperal oxygen saturation
- PaCO2 - Arterial partial pressure of carbon dioxide
- pH - Arterial pH
- BE - Arterial base excess
- Lactate - Arterial lactate
- NAdose - Noradrenaline dose
- MAP - Mean arterial pressure
- Creatinine -
- Urea -
- CK - Creatinine Kinase
- Bilirubin -
- Albumin -
- CRP - C reactive protein
- Fibrinogen -
- Ddimer -
- ATIII - Anti-thrombin III
- HB - Haemaglobin
- Leukocytes -
- Platelets -
- TNFa -
- IL6 -
- IL8 -
- siL2

20.3 PreECMO data

Question(s) of interest

The main questions of interest are:

- 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?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.

20.4 Day1ECMO data

Question(s) of interest

The main questions of interest are:

- Can we use the Day1ECMO biomedical markers to accurately predict ECMO survival?
- Do we need all Day1ECMO variables or just a subset to make accurate predictions?
- What is our expected future performance for these predictions?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.

21 Bayesian way of predicting the price of gold (1)

Project level: Difficult

21.1 Overall project description

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.

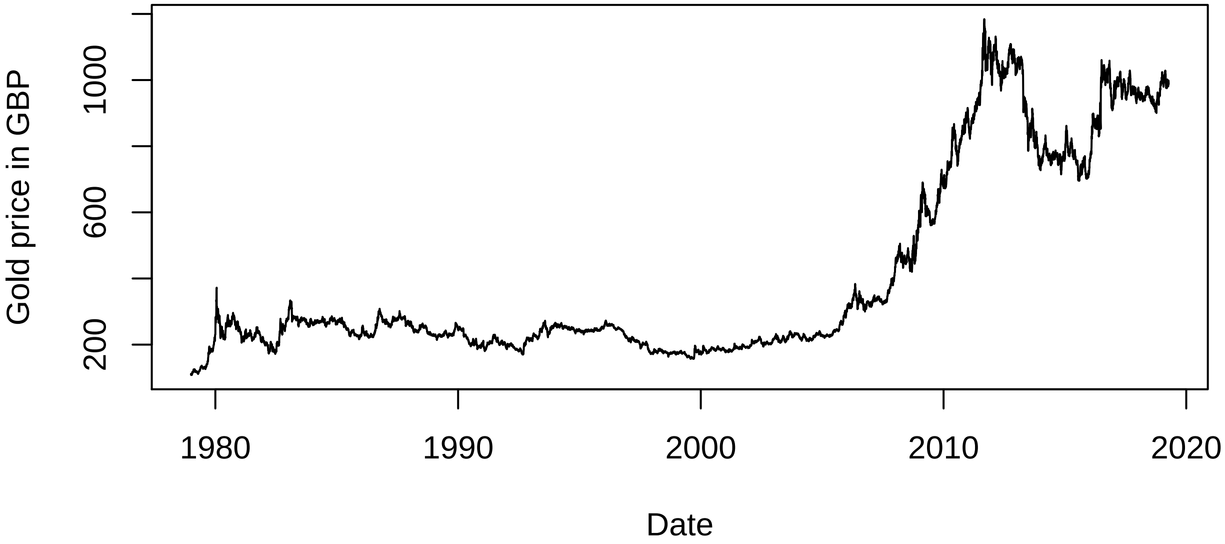
To tackle the problem of likelihood intractability, we will be using the Approximate Bayesian Computation (ABC) methods for approximate inference and prediction.

21.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Historical gold prices can be obtained from http://gold.org for research purposes. The complete data set contains many different price summaries, we will be using daily average prices in Pounds Sterling from 1978 until early 2019.



The Black-Scholes model assumes a simple linear model for the average price change, while the volatility of the price is considered to be stochastic. The SDE describing the price S of the commodity is defined as the following:

$$\frac{dS}{S} = \mu dt + \sigma dW$$

where, in our case, S is the price of gold, W is a stochastic variable (Brownian motion). Note that W, and consequently its infinitesimal increment dW, represents the only source of stochasticity in the price history. Intuitively, $W(t)$ is a process that “wiggles up and down’ ’ in such a random way that its expected change over time is zero. In addition, its variance over time T is equal to T. The parameters of this model, μ and σ, define the rate of average price change and its volatility, correspondingly.

We will treat this problem in a Bayesian way. We will, therefore, assign some weakly informative priors to the unknown parameters μ and σ, perform approximate inference of the posteriors of these parameters given historical prices, and finally produce posterior predictive distributions for possible gold prices within the next month.

Main approach for inference will be using the Approximate Bayesian Computation methods that rely on simulating samples from the SDE model (using the Euler–Maruyama method) and comparing these samples to the observed data. However, an explicit solution actually exists for the Black-Scholes model. It can be demonstrated that:

$$S(t) = S(0) \exp\left\{\sigma W_t + \left(\mu - \frac{1}{2}\sigma^2\right) t\right\}$$

It might be interesting to perform inference and prediction using this explicit solution, and observing how large is the approximation error when using ABC methods.

Question(s) of interest
The main questions of interest are:

- Using appropriate priors can inference of the model parameters be done when working with SDE?
- Can predictions be made using the information learned from historical data?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Bayesian Statistics.
- Advanced Bayesian Methods.

22 Bayesian Linear Models and Bayesian Lasso (4)

Project level: Difficult

22.1 Overall project description

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.

A number of projects are available on this topic considering different data sets.

22.2 Individual project details

How many individual projects are available in this area: 4.

22.3 Communities and Crime Data Set

Data available
The data set available from the following address: http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

The data set contains 1994 records of 128 variables. The response variable is `ViolentCrimesPerPop`, total number of violent crimes per 100K population. The other variables describe different demographical characteristics of US neighbourhoods. Your goal is to build a linear model for predicting the rate of violent crimes from neighbourhood characteristics.

First, use a Bayesian formulation of a linear model to infer regression rates. Next, perform Bayesian Lasso for variable selection to decide what are the most informative factors to explain crime rate.

Question(s) of interest
The main questions of interest are:

- Can a linear model be formulated to perform inference in a Bayesian framework?
- Can variable selection using Bayesian Lasso determine the most important factors in the variation in crime rates among neighbourhoods?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Bayesian Statistics.
- Advanced Bayesian Methods.

22.4 Concrete Slump Test Data Set

Data available

The data set available from the following address: http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test.

The data set contains 103 records of 10 variables. The response variable is `28-day Compressive Strength (Mpa)`, the compressive strength measure of concrete slab. Only 7 of the variables should be considered as explanatory variables, these are proportions of different components in concrete measured in kg per m^3:

- Cement
- Slag
- Fly ash
- Water
- SP
- Coarse Aggr.
- Fine Aggr.

To investigate interaction between explanatory variables, you may want to introduce latent explanatory variables, such as pairwise products of the above component concentrations, or their ratios.

First, use a Bayesian formulation of a linear model to infer regression rates. Next, perform Bayesian Lasso for variable selection to decide what are the most informative factors to explain strength of the resulting concrete.

Question(s) of interest

The main questions of interest are:

- How to formulate a linear model in the Bayesian framework?
- How to perform conjugate and non-conjugate inference for Bayesian linear model?
- How to perform variable selection using Bayesian Lasso?
- What are the most important factors to explain strength of a concerete slab?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Bayesian Statistics.
- Advanced Bayesian Methods.

22.5 Wine Quality Data Set

Data available

The data set available from the following address: http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

The data set contains 4898 records of 12 variables. The response variable is `quality`, the score for the quality of a particular wine. The rest of the variables are explanatory variables:

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

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.

First, use a Bayesian formulation of a linear model to infer regression rates. Next, perform Bayesian Lasso for variable selection to decide what are the most important factors that make a good wine.

Question(s) of interest
The main questions of interest are:

- How to formulate a linear model in the Bayesian framework?
- How to perform conjugate and non-conjugate inference for Bayesian linear model?
- How to perform variable selection using Bayesian Lasso?
- What are the most important factors to make a good wine?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Bayesian Statistics.
- Advanced Bayesian Methods.

22.6 Boston Housing Data Set

Data available
The data set available from `mlbench` package in `R`. Make sure you install this package first. You can load the data using the following code:

```
require(mlbench)
```

```
## Loading required package: mlbench
```

```
data(BostonHousing)
d <- BostonHousing
y <- d$medv
X <- d[,-14]
```

The data set contains 506 records of 14 variables. The response variable is `MEDV` that represents the median value of the owner-occupied homes in the census tract for different neighbourhoods in Boston in 1970s. The rest of the variables are explanatory variables.

- RM - Number of rooms in owner units
- AGE - Proportion of units built prior to 1940
- B - Racial mix
- LSTAT - Percentage of low income earners
- CRIM - Crime rate by town
- ZN - Proportion of residential area zoned for large lots
- INDUS Proportion of non-retail business acres per town
- TAX Full value property tax rate
- PTRATIO Pupil to teacher ratio
- CHAS Charles River location
- DIS Weighted distances to five employment centres
- RAD Accessibility to highways
- NOX Nitric oxide concentration

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.

First, use a Bayesian formulation of a linear model to infer regression rates. Next, perform Bayesian Lasso for variable selection to decide what are the most important factors that define real estate value.

Question(s) of interest
The main questions of interest are:

- How to formulate a linear model in the Bayesian framework?
- How to perform conjugate and non-conjugate inference for Bayesian linear model?
- How to perform variable selection using Bayesian Lasso?
- What are the most important factors for the price of real estate?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Bayesian Statistics.
- Advanced Bayesian Methods.

23 Analysis of Course and Degree Results in a University Programme (3)

Project level: Easy

23.1 Overall project description

University undergraduate students are often interested in knowing which combinations of courses are most likely to lead to the award of the highest class of degree.

Data are available from a university statistics department on course choices and course results for students (`Sudent.Num`) studying statistics courses in their second, third and final years (`Levels` "2", "3" and "4" in the data) for three cohorts of students: namely those graduating in 2016, 2017 and 2018, respectively. The final class of degree awarded is also available (`Degree.Classification`), as is a description of the degree programme (`Programme`) that each student was enrolled on, namely:

- `Single` - Single Honours in Statistics
- `Maths and Stats` - Combined Honours in Statistics & Mathematics
- `Stats & Other` - Combined Honours in Statistics & a non-Maths Subject
- `Erasmus` - Statistics as part of a EU Programme

Course results are listed under/besides their course codes as primary and secondary grades on a 22 point scale, as illustrated in Table 2.2 here https://www.gla.ac.uk/media/media_124293_en.pdf. In addition to these grades, a medical or other adverse circumstances exemption grade of "MV" or a credit withheld grade of "CW" or a credit refused grade of "CR" may be awarded when the student failed to comply, in the absence of good cause, with the published requirements of the course or programme.

For the 3rd and 4th year results, the total number of credits (`Credits.Level3` and `Credits.Level4`, respectively) out of a fulltime study total of 120 are given for each student. Note that each course is worth 10 credits with the exception of `SProj.30Cr` and `SProj.20Cr` which refer to final year projects worth 30 and 20 credits, respectively.

The course results are combined in 3rd and 4th years to produce the aggregate scores `Level3Aggregate` and `Level4Aggregate`, respectively, and then combined together to produce `Overall.Aggregate`. These aggregate scores are used to produce the `Degree.Classification` on a 5 point scale, as defined in Table 2.3 here https://www.gla.ac.uk/media/media_124293_en.pdf.

23.2 Individual project details

How many individual projects are available in this area: 3.

23.3 Relationships between choices of courses in final year and class of Honours Degree

This project focuses on the course choices and results in the final year of study.

Question(s) of interest
The main questions of interest are (you are not limited to exploring these):

- Are there differences in the distribution of course results?
- Are there differences in the course and degree results between the different programmes of study?
- Are there any differences in the course and degree results between years of graduation?
- What is the relationship between the aggregate result from third year statistics courses and the overall aggregate score and class of degree awarded.
- What is the relationship between the choice of final year statistics courses and the overall aggregate score and class of degree awarded? If there is a relationship, which courses should be chosen in the final year of study to maximize the chances of a first class degree?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.
- Statistical Inference.
- Generalised Linear Models.

23.4 Relationships between choices of courses in third year of study and class of Honours Degree

This project focuses on the course choices and results in the third year of study.

Question(s) of interest
The main questions of interest are (you are not limited to exploring these):

- Are there differences in the distribution of course results?
- Are there differences in the course and degree results between the different programmes of study?
- Are there any differences in the course and degree results between years of graduation?
- What is the relationship between the aggregate result from third year statistics courses and the overall aggregate score and class of degree awarded.
- What is the relationship between the choice of third year statistics courses and the overall aggregate score and class of degree awarded? If there is a relationship, which courses should be chosen in the final year of study to maximize the chances of a first class degree?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.

- Statistical Inference.
- Generalised Linear Models.

23.5 Relationships between performance in second year and class of Honours Degree

This project focuses on the course results in the second year of study.

Question(s) of interest
The main questions of interest are (you are not limited to exploring these):

- Are there differences in the distribution of course results?
- Are there differences in the course and degree results between the different programmes of study?
- Are there any differences in the course and degree results between years of graduation?
- Is there a relationship between a students performance in second year and their aggregate scores in 3rd and 4th year and their overall aggregate score and class of degree awarded?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.
- Statistical Inference.
- Generalised Linear Models.

24 Optical character recognition (1)

Project level: Moderate

24.1 Overall project description

Image processing is a difficult task for machines. The relationships linking patterns of pixels to higher concepts are complex and hard to define. For instance, it is easy for a human being to recognise a face or a letter, but defining these patterns in strict rules is difficult. Furthermore, image data are often noisy. There can be many slight variations in how the image was captured depending on the lighting, orientation and positioning of the subject. This project in particular is about optical character recognition (OCR), where the objective is to differentiate among the 26 letters of the English alphabet based on handwritten letters, like shown in the following image:

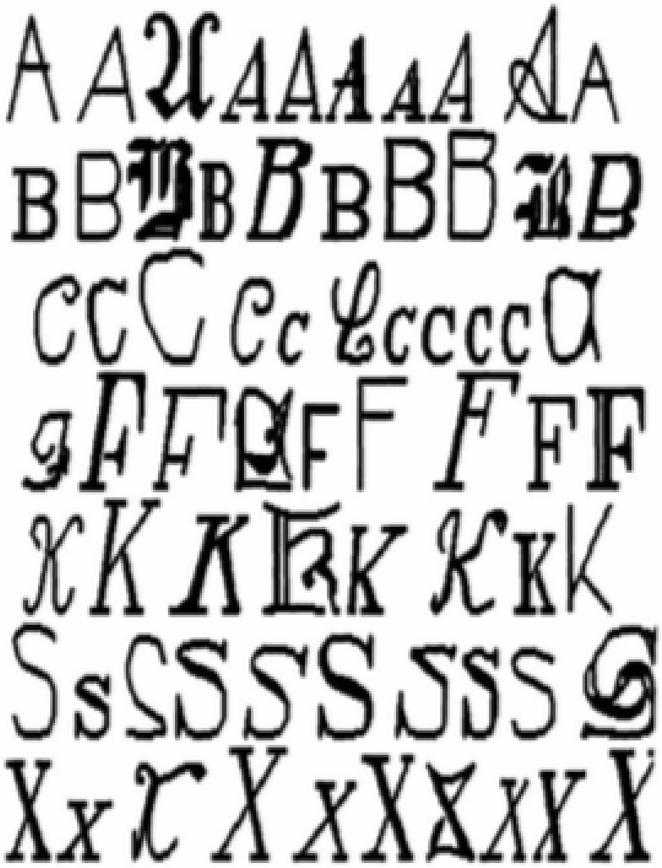


Figure 1. Examples of the character images generated by "warping" parameters.

For the following project, 20,000 handwritten characters were scanned into a computer, converted into pixels and 16 statistical attributes were recorded, following a procedure proposed by Frey and Slate. These attributes measure such characteristics as the horizontal and vertical dimensions of the letter, the proportion of black versus white pixels, and the average horizontal and vertical position of the pixel. The task of this project is to develop and assess a classifier that reads in these attributes and predicts the letter.

24.2 Individual project details

How many individual projects are available in this area: 1.

Data available
The data are available in file `letterdata.txt`. The first line is a standard line of headings, where the first column (y) indicates the letter, and the following 16 columns (x01 to x16) are 16 integer numbers with the attributes mentioned above.

Question(s) of interest
What classification performance can be obtained with a statistical method, i.e. how close can a machine using a statistical pattern recognition algorithm get to human performance? How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

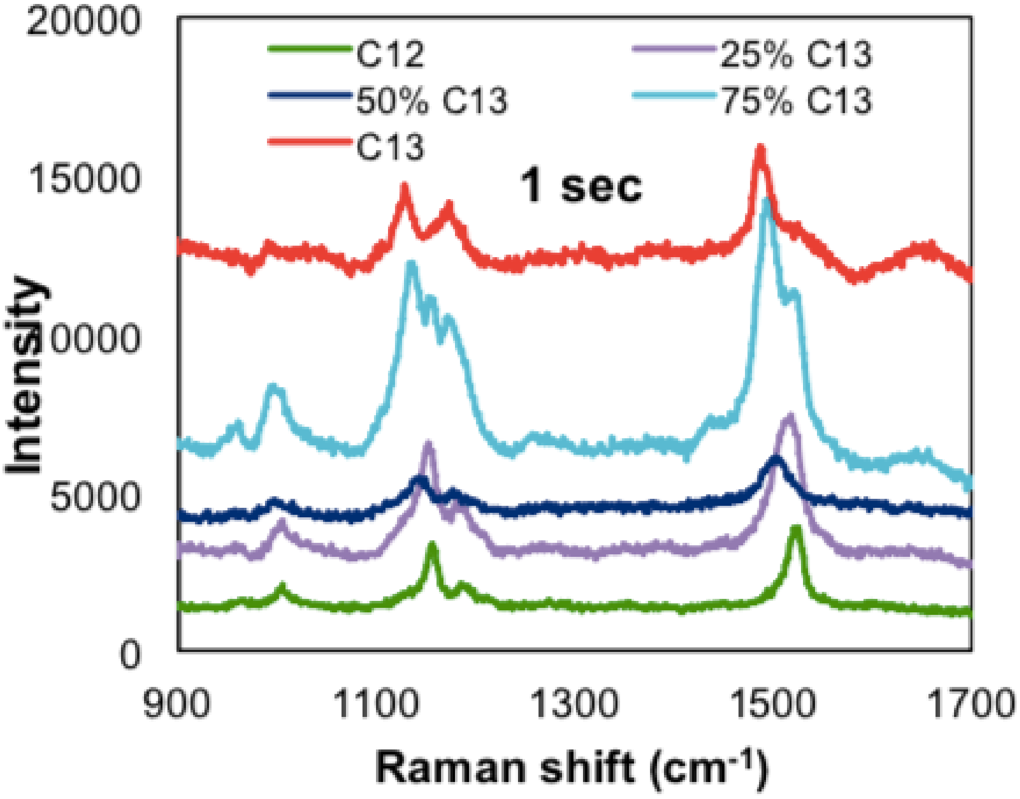
- Inference.
- Flexible regression.
- Generalised linear models.
- Mutivariate methods.
- Big data analytics.
- Introduction to R programming.

25 Classifying bacterial metabolic states with Raman spectroscopy (1)

Project level: Moderate

25.1 Overall project description

Raman spectroscopy is a spectroscopic technique used to observe low-frequency modes in a molecular system and is commonly used in chemistry to provide a structural fingerprint by which molecules can be identified. It relies on inelastic scattering of monochromatic light, usually from a laser in the visible, near infrared, or near ultraviolet range. The laser light interacts with molecular vibrations, phonons or other excitations in the system, resulting in the energy of the laser photons being shifted up or down. The shift in energy gives information about the vibrational modes in the system. A set of typical Raman spectra is shown in the figure below.



The objective of the present project is to distinguish between different metabolic states in two unicellular organisms: Chlorella (a single-celled green algae), and Rhodobacter (a proteobacterium). The Raman spectra were obtained in Professor Huabing Yin's group in the School of Engineering, and include 171 strains of Chlorella, and 139 strains of Rhodobacterium. The spectra are discretized, and show the normalized scatter intensities at 498 discrete laser wavelengths. For both unicellular organisms, there are 5 different metabolic states. The objective is to build a statistical classifier to correctly predict the metabolic state from the Raman spectra. To this end, you want to develop and assess a range of classifiers that read in the Raman spectra and predict the metabolic state of the unicellular organism.

25.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data are available in the files `data_chlorella.txt` and `data_Rhodo.txt`. The first line is a standard line of headings, where the first column (y) indicates the metabolic state (an integer number between a and 5), and the following 498 columns (x001 to x498) show the standardized scatter intensities at 498 laser wavelengths.

Question(s) of interest

How accurately can we predict bacterial metabolic states from Raman spectra? In other words, can we use the set of 498 discrete x values to predict the metabolic states in each organism? How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Inference.
- Flexible regression.
- Generalised linear models.
- Multivariate methods.
- Big data analytics.
- Introduction to R programming.

26 Reading attainment in primary school children (1)

Project level: Moderate

26.1 Overall project description

You are given a data set that arose from a longitudinal study of a cohort of 407 pupils in 33 multi-ethnic infant schools in London. The reading ability of the pupils was tested on up to six occasions: annually over five years, starting with the year when they entered the school, and 3 years later at the end of their junior schooling. Data are also available on the age of the pupils at the occasions when the testing was performed and also their gender and ethnic group. The pupils took a variable number of assessments and so the data are unbalanced. The data are contained in the file `reading.dat`, which has eight columns:

- School number (1 to 33)
- Pupil number (1 to 751)
- Assessment occasion (1 to 6)
- Reading attainment score
- A standardisation score that can be ignored
- Ethnic group
- Gender (boy or girl)
- Age (in years, but mean-centred)

26.2 Individual project details

How many individual projects are available in this area: 1.

Question(s) of interest
Some questions of interest are:

- How does reading ability develop as children grow older?
- Does this ability vary from pupil to pupil or from school to school?
- If so, does it vary systematically from one type of pupil to another (e.g. boys versus girls, white versus black, or both?)

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Linear mixed models.
- Inference.
- Introduction to R programming.

27 Finding epigenetic signatures for human ageing (2)

Project level: Moderate/Difficult

27.1 Overall project description

Biological ageing of human cells is one of the primary risk factors for the development of cancer or other lethal diseases. The biology of ageing is a complex process, involving many layers of interactions among the components of the human cell. Actual chronological age may not be a good measure for biological age as people may age at different rates, due to genetic, environmental or even lifestyle factors. However, recently developed laboratory experiments allow for the measurement of various biological factors, from clinical-level measurements to epigenetic ones, such as alterations in the chromosome (histone modifications) and methylation of the DNA, that affect gene activity and function and impact ageing of cells. In this project, you will analyse epigenetic data, on histone modifications and methylation at sites in human DNA, in proliferating ("young") and senescent ("old") human cells, to determine a characterization (signature) for biological ageing and estimate the effects of these factors on the human ageing process.

27.2 Individual project details

How many individual projects are available in this area: 2.

27.3 Stratification of ageing-associated modifications in human DNA

Data available
Data on several histone modifications and methylation, measured on about 2100 ageing-associated CpG sites in human DNA, from proliferating ("young") and senescent ("old") human cells is available. These sites have been determined, through other biological studies, to be ageing-associated differentially methylated positions (aDMPs) in the DNA. The data are stored in `aDMPs_Proj1.csv` and contain the following columns.

- Column 1: `CpG_ID` - A unique code for each site
- Columns 2-7: `Prolif_H3.3`,`Prolif_H4K16ac_ab1`,`Prolif_H4K16ac_ab2`,

  `Prolif_H4K20me3_ab1`,`Prolif_H4K20me3_ab2`,`Prolif_H4`
  - Abundance of 6 different types of histone modification at each CpG site, in proliferating ("young") cells
- Column 8: `Prolif_Meth` - Methylation ratio at each CpG site, in proliferating ("young") cells

- Columns 9-14: \texttt{Senes_H3.3,Senes_H4K16ac_ab1,Senes_H4K16ac_ab2,

  `Senes_H4K20me3_ab1,Senes_H4K20me3_ab2,Senes_H4`

  - Abundance of 6 different types of histone modification at each CpG site, in senescent ("old") cells}

- Column 15: `Senes_Meth` - Methylation ratio at each CpG site, in senescent ("old") cells

- Column 16: `Correlation.with.Age` - Spearman Correlation Coefficient for how well a CpG site's level of methylation correlates with biological age.

Question(s) of interest
The main questions of interest are:

- Can the aDMPs be stratified based on the measured abundances of histone modifications and methylation observations? Does the stratification vary across proliferating cells, senescent cells, or both types of cells taken together?
- What is the effect of each histone modification on the propensity for cell ageing?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression modelling.
- Multivariate methods.
- Machine Learning.

27.4 Determining a epigenetic signature for biological ageing

Data available
Data on several histone modifications and methylation, measured on about 285,000 CpG sites in human DNA, from proliferating ("young") and senescent ("old") human cells is available, measured from an Illumina 450k methylation array. The data are stored in `aDMPs_Proj2.csv` and contain the following columns.

- Column 1: `CpG_ID` - A unique code for each site

- Columns 2-7: `Prolif_H3.3,Prolif_H4K16ac_ab1,Prolif_H4K16ac_ab2`,

`Prolif_H4K20me3_ab1,Prolif_H4K20me3_ab2,Prolif_H4`

- Abundance of 6 different types of histone modification at each CpG site, in proliferating ("young") cells

- Column 8: `Prolif_Meth` - Methylation ratio at each CpG site, in proliferating ("young") cells

- Columns 9-14: `Senes_H3.3,Senes_H4K16ac_ab1,Senes_H4K16ac_ab2`,

`Senes_H4K20me3_ab1,Senes_H4K20me3_ab2,Senes_H4`

- Abundance of 6 different types of histone modification at each CpG site, in senescent ("old") cells}
- Column 15: `Senes_Meth` - Methylation ratio at each CpG site, in senescent ("old") cells
- Column 16: `aDMP_status` - Binary variable indicating whether a CpG site is an ageing-associated differentially methylated position or aDMP (1) or not (0).

Question(s) of interest
The main questions of interest are:

- Can aDMPs be distinguished from the non-aDMPs based on the histone abundance and methylation observations (i.e. is there an epigenetic signature for aDMPs)?
- Does the epigenetic signature exist, or vary, across proliferating cells, senescent cells, or both types of cells taken together?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Multivariate methods.
- Machine Learning.
- Bayesian statistics.

28 Determining genetic variation associated with heart disease (2)

Project level: Moderate

28.1 Overall project description

It has long been known to scientists and clinicians that heart disease is a complex set of conditions that are caused in part by environmental or lifestyle factors, but also has a significant connection to the underlying genetics of an individual. The genetic signature of every human being is unique, encoded in their DNA, which can be represented as a long (of length about 3 billion) string of nucleotides, A, C, G and T, in some specific order. Many common health conditions are caused due to variations from the "normal" DNA at a few specific positions on the genome. Genetic variation in individuals often occurs as single alterations (mutations) in different positions of the genome, termed "single nucleotide polymorphisms" or SNPs. Genome-wide association studies (GWAS) are a popular method for studying and determining the locations of these SNPs. Using experimental plates that contain millions of SNPs from hundreds or thousands of people, the goal of GWAS is to detect which SNPs are associated with a particular disease outcome. Much recent evidence indicates that two or more SNPs often work in combination to produce a genetic effect, which suggests that multiple regression methods with variable selection may be a potential way to determine causal SNPs.

In this project, you will study genetic and clinical data collected at a Glasgow medical centre and try to determine which factors play a part in the development of heart disease. The *genotype*, or genetic composition at each location of the genome is typically given by one of 3 possibilities, aa, ab, or bb, where a and b take values from the set {A,C,G,T}. These three possibilities are usually encoded numerically by 0, 1, and 2 for purposes of statistical analysis. In a typical genetic experiment (called a genome-wide association study) to study the effect of genetic variation on some characteristic (*phenotype*) or disease, data is collected from thousands of individuals with varying levels of the phenotype (or disease), and their DNA sequenced for about 500,000-1,000,000 locations on their genomes. It is still an extremely challenging problem to detect which SNPs are associated with the phenotype of interest, compounded by high volumes of data, high levels of missingness, and high correlations among SNPs that are located in certain neighborhoods in the genome.

In this project, you will study a simplified version of this problem in which a small set of candidate SNPs (that have been selected by other means, and may have an impact on the phenotype of interest) are given to you, along with a number of measurements on certain clinical covariates, and measurements on some features representing aspects of heart disease.

28.2 Individual project details

How many individual projects are available in this area: 2.

28.3 Determining genetic factors associated with high blood pressure

Data available

Data is provided in two files. The first file `bloodpressure.csv`, contains information on systolic and diastolic blood pressure for patients at the clinic, along with a number of clinical measurements. The file contains the following columns.

- Column 1: `IID` - A unique code for each individual
- Column 2: `age`
- Column 3: `sex` (1: male; 2: female)
- Column 4: `bmi` (body mass index)
- Column 5: `newsmoke` (1: if person has started smoking; 0: non-smoker)
- Column 6: `sbp` (systolic blood pressure)
- Column 7: `dbp` (diastolic blood pressure)

The second file, `snpdata.csv` contains information on the candidate SNPs for each individual. There are 14 SNPs in total, with the unique SNP id given in the column header. The columns of the file are:

- Column 1: `IID` - A unique code for each individual in the study (there are fewer individuals in this file than in the file containing the blood pressure measurements)
- Column 2: `sex` (1: male; 2: female)
- Columns 3-16: value of the SNP at the measured location. SNPs take values of 0, 1 or 2, according to how many instances of the minor allele (less prevalent nucleotide) are present at that location.

Question(s) of interest

The main questions of interest are:

- Are any of the measured clinical covariates associated with high blood pressure?
- Do one or more of the candidate SNPs appear to be associated with high blood pressure?
- How much of blood pressure variation can be explained by clinical/lifestyle factors, genetic factors, or both?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression modelling.
- Data Analysis.
- Multivariate methods.
- Machine Learning.
- Big Data Analytics.
- Bayesian Statistics.
- Advanced Bayesian methods.

28.4 Determining genetic and lifestyle factors underlying blood sugar and cholesterol levels

Data available

Data on several clinical measurements and candidate SNPs are available for this project, stored in the file `gwasHDLglu.csv`, containing the following columns:

- Column 1: `IID` - A unique ID for each individual in the study
- Column 2: `age`
- Column 3: `sex` (1: male; 2: female)
- Column 4: `bmi` (body mass index)
- Column 5: `newsmoke` (1: if person has started smoking; 0: non-smoker)
- Column 6: `prevcvd` (incidence of previous cardiovascular disease; 1 if true)
- Column 7: `Fglu` (fasting glucose level)
- Column 8: `HDL` (high-density lipoprotein or “good” cholesterol level)
- Columns 9-22: value of the SNP at the measured location. SNPs take values of 0, 1 or 2, according to how many instances of the minor allele (less prevalent nucleotide) are present at that location.

Question(s) of interest

The main questions of interest are:

- How do levels of fasting glucose and HDL vary among different segments of the clinical population?
- How well can the variation in fasting glucose be explained by lifestyle factors?
- Can fasting glucose level prediction be improved by accounting for genetic variation in specific SNPs?
- Are HDL levels associated with lifestyle factors, genetic factors, or both?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression modelling.

- Data Analysis.
- Multivariate methods.
- Machine Learning.
- Big Data Analytics.
- Bayesian Statistics.
- Advanced Bayesian methods.

29 Statistical analysis of US presidential election data (2)

Project level: Moderate

29.1 Overall project description

In the two recent US Elections, there have been much speculation on whether various socio-economic and demographic factors, such as race, age and income levels (to name a few) played a role in the preference for political parties or candidates. Presidential elections in the USA occur every four years, with registered voters casting their ballots on Election Day, which is the first Tuesday after November 1 that year. The modern political system in the U.S. is a two-party system dominated by the Democratic Party and the Republican Party. These two parties have won every United States presidential election since 1852, alternating on a fairly regular basis. In this set of projects, you will be asked to make use of presidential election data and demographic data from the US Census Bureau to analyse potential associations between socio-economic-demographic groups and electoral results in various states and counties in the USA.

29.2 Individual project details

How many individual projects are available in this area: 2.

29.3 Analysis of 2012 Presidential election data

The data set provided, `election2012.csv`, gives a number of demographic characteristics for each state (from the US Census Bureau web site, http://www.census.gov), along with the electoral outcomes in that state, for the 2012 Presidential election. The variables are listed in the following order:

- Column 1: `State` (name of State)
- Column 2: `State.ID` (2-letter ID for state)
- Column 3: `won` (which party won D- democratic; R- Republican)
- Column 4: `Sep12unempl` (Percent unemployed in September 2012)
- Column 5: `Unempl.changeJan09` (Change in percent unemployed between Jan 2009 and Sep 2012)
- Column 6: `PercPoverty` (Percent of population in poverty)
- Column 7: `UrbanPop2000` (Percent of population living in urban areas)
- Column 8: `Over65` (Percent of population aged 65 or higher)

- Column 9: `PercFemale` (Percentage of female population)
- Column 10: `High.school.or.less` (Percent who have a high school degree or less)
- Column 11: `Graduate.deg` (percent having graduate or professional degrees)
- Column 12: `No.health.insurance` (percent with no health insurance)
- Column 13: `African.American` (Percent African American or Black)
- Column 14: `Hispanic` (Percent Hispanic or Latino)

Question(s) of interest
Using this data, you will try to assess whether various demographic characteristics seem to have a possible influence on electoral outcome. In particular, the main questions of interest are:

- Are there any particular groups or clusters of states characterised by certain patterns of socio-economic and demographic factors?
- Are there specific combinations of social and/or economic factors that tend to favour either the Democratic or Republican party winning in a state?
- Which states seemed most important in decided the outcome of the 2012 election, and why?
- Can socio-economic-demographic factors alone be used to predict the electoral outcome in states? How well can the election results be predicted from these?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Data Analysis.
- Multivariate methods.
- Machine Learning.
- Big Data Analytics.

29.4 Assessing the impact of socio-economic factors on Presidential primary election voting in the USA in 2016

Data available
You are provided a data set on Presidential election results for each US county in 2016 `PresElect2016R.csv` and socio-economic data from the US Census Bureau (until 2014), in the file `UScounty-facts.csv`. An additional file, `UScounty-dictionary.csv`, is provided, which lists the detailed descriptions of variables available in the county facts file. For the purposes of this analysis, you may assume that there was an election involving only two parties in each county: Republican and Democratic. A brief description of the variables in the files are listed below:

File 1: `PresElect2016R.csv`

- Column 1: `state`

- Column 2: `state.po` (2-letter state abbreviation)
- Column 3: `county` (county name)
- Column 4: `FIPS` (unique ID for county from US Census records)
- Column 5: `candidatevotesR` (number of votes cast for Republican presidential candidate)
- Column 6: `totalvotes` (total number of votes cast in the county)
- Column 7: `fracvotesR` (fraction of total votes received by the Republican Presidential candidate)
- Column 8: `partywonR` (binary variable that takes the value 1 if the Republican candidate won in that county; is otherwise zero)

File 2: `UScounty-facts.csv`

The columns of this file correspond to measurements on several variables for each county, described in `UScounty-dictionary.csv`. Variables 1-18 correspond to demographic variables relating to the population and racial composition of counties. Variables 19 and 20 correspond to educational attainment; variable 21 to the number of war veterans in the county; variables 22-28 relate to housing; variables 29-42 to income and employment; variables 43-47 to sales; and variables 48-50 to building permits, land area and population per square mile, respectively.

Question(s) of interest

The main questions of interest are twofold: first, are there any discernible associations between various socio-economic and other factors and the propensity of the county population to vote for a particular party? Second, can the relationship between various factors and primary election results by county be consolidated into a model that can forecast the actual 2016 presidential election results, by state? In particular, you may want to consider:

- Are there specific socio-economic or demographic factors that are associated with an increased or decreased preference for a political party, in a county?
- Is there an association between specific socio-economic or demographic factors and the fraction of people voting for a Republican Presidential candidate in a county?
- Are there state-wide factors that are associated with a preference for one political party over another?
- How well can your model associating socioeconomic factors with 2016 election results be used to predict the final state-wide outcome of the presidential elections in 2016? (For this question you might want to locate a data set listing the winning party in each state- this is available on numerous internet news sites, such as CNN.com or NPR.org; alternatively, you can consolidate data from within your existing data set.)

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Regression models.

- Data Analysis.
- Multivariate methods.
- Machine Learning.
- Big Data Analytics.
- Bayesian Statistics.

30 Predicting Olympic medal counts (1)

Project level: Moderate

30.1 Overall project description

The aim of this project is to develop models for predicting the number of medals won by each country at the Rio Olympics in 2016 using information that was available prior to the Games. The emphasis is on prediction, so appropriate measures should be used to evaluate the predictive performance of models used. Finally a comparison should be made to some of the predicted rankings/medal counts published online just before the Olympics in 2016.

30.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on the number of medals (total and gold) won by each country for 108 countries participating in the Rio 2016 Olympics, along with information on previous Olympic performance (from the 2000, 2004, 2008 and 2012 Games) and other variables.

It is also possible to augment the data by adding variables to the list below, provided that these variables were available before the beginning of the Games in August 2016.

The dataset `olympics2016.csv` has 108 observations and the following variables:

- `country` the country's name,
- `country.code` the country's three-letter code,
- `gdpYY` the country's GPD in millions of US dollars during year YY,
- `popYY` the country's population in thousands in year YY,
- `soviet` 1 if the country was part of the former Soviet Union, 0 otherwise,
- `comm` 1 if the country is a former/current communist state, 0 otherwise,
- `muslim` 1 if the country is a Muslim majority country, 0 otherwise,
- `oneparty` 1 if the country is a one-party state, 0 otherwise,
- `goldYY` number of gold medals won in the YY Olympics,
- `totYY` total number of medals won in the YY Olympics,
- `totgoldYY` overall total number of gold medals awarded in the YY Olympics,
- `totmedalsYY` overall total number of all medals awarded in the YY Olympics,

- `bmi` average BMI (not differentiating by gender),
- `altitude` altitude of the country's capital city,
- `athletesYY` number of athletes representing the country in the YY Olympics,
- `host` 1 if the country has hosted/is hosting/will be hosting the Olympics, 0 otherwise.

The first observation, corresponding to Afghanistan, is shown below.

```
oldat <- read.csv("olympics2016.csv")
oldat$gdp16 <- as.numeric(oldat$gdp16)
```

```
## Warning: NAs introduced by coercion
```

```
head(oldat,1)
```

```
##       country country.code gdp00 gdp04 gdp08 gdp12 gdp16 pop00 pop04 pop08
## 1 Afghanistan          AFG  #N/A  5285 10191 20537 19469 20094 24119 27294
##   pop12 pop16 soviet comm muslim oneparty gold00 gold04 gold08 gold12 gold16
## 1 30697 34656      0    0      2        0      0      0      0      0      0
##   tot00 tot04 tot08 tot12 tot16 totgold00 totgold04 totgold08 totgold12
## 1     0     0     1     1     0       298       301       301       301
##   totgold16 totmedals00 totmedals04 totmedals08 totmedals12 totmedals16  bmi
## 1       298         915         924         949         956         949 23.3
##   altitude athletes00 athletes04 athletes08 athletes12 athletes16 host
## 1     1790          0          5          4          6          3    0
```

Question(s) of interest

The main questions of interest are:

- Which variables are associated with the number of medals (total/gold or both) won in the 2012 Olympics?
- How well does a model based on data up to and including 2012 predict Olympic performance in the 2016 Games?
- What improvements might be made to the model/data collected in order to better predict Olympic medal counts for future Games?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Generalised Linear Models.
- Linear Mixed Models.
- Flexible Regression.

31 Monitoring physical and mental health outcomes in Scotland (2)

Project level: Moderate

31.1 Overall project description

The Scottish Health Survey monitors the health of the Scottish population living in private households. The main aim of the survey is to keep an eye on health trends in Scotland. Data from the Scottish Heath Surveys from 2008 to 2016 will be used to explore physical and mental health outcomes as a function of demographic, socioeconomic and lifestyle factors. The project will also focus on how Glasgow compares to other parts of Scotland to examine whether a "Glasgow effect" remains after adjusting for the above factors.

31.2 Individual project details

How many individual projects are available in this area: 2.

Data available

Data are available on health outcomes, socio-economic and lifestyle factors from the 2008-2016 Scottish Health Surveys. The data are stored in `shs.Rdata`. The R object `shs` contains the following columns:

- `Age` - Age of individual
- `Agegroup` - "16-24", "25-34", ..., "75+"
- `Sex` - Male or Female
- `Smoking` - Current cigarette smoker, Ex-smoker, or Never smoked
- `Education` - Highest educational qualification of individual
- `Birthplace` - Elsewhere, England, Wales or Northern Ireland or Scotland
- `Alcohol` - Drinks outwith government guidelines, Drinks within government guidelines, Ex drinker or Never drank alcohol
- `Employment` - Doing something else, In full-time education, In paid employment, self-employed or on gov't training, Looking after home/family, Looking for/intending to look for paid work, Perm unable to work, Retired
- `CMOrec` - Chief Medical Office guidance for physical activity: Meets muscle rec only, Meets MVPA & muscle recs, Meets MVPA and muscle recs, Meets MVPA rec only, Meets neither rec
- `Veg` - Consume recommended daily vegetable intake (Yes/No)
- `Fruit` - Consume recommended daily fruit intake (Yes/No)
- `HealthBoard` - Scottish Health Board (18 total)
- `Longillness` - Long-term illness (Yes/No)
- `SAgenHealth` - Self-assessed general health, Fair/bad/very bad or Very good/good

- `GHQ` - General health questionnaire (score between 0 and 12), high values indicate possible psychiatric disorders
- `WEMWBS` - Warwick-Edinburgh Mental Well-Being Scale (score between 14 and 70), higher scores indicate higher positive mental well-being
- `Cardio` - Cardiovascular condition (Yes/No)
- `Lifesat` - Life satisfaction (below or above the mode)
- `BP` - High blood pressure (Yes/No)
- `BMI` - Body Mass Index of individual
- `Year` - Year of the Scottish Health Survey
- `Glasgow` - Health board in Glasgow area (Yes/No)
- `BMIgroup` - Normal, Obese, Overweight, Underweight

31.3 Physical health in Scotland 2008-16

Questions of interest

The main questions of interest are:

- Which variables/factors are associated with physical health outcomes such as having a cardiovascular condition, high blood pressure and self-assessed general health?
- Are there any trends or changes in patterns of physical health as described by the above variables over the years of the Scottish Health Survey?
- Are there any differences in the above physical health outcomes between Glasgow and the rest of Scotland, after adjusting for demographic, socioeconomic and lifestyle factors?
- What other variables might be useful in better understanding what influences physical health in Scotland?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised Linear Models.
- Regression Modelling.
- Flexible Regression.

31.4 Mental health in Scotland 2008-16

Questions of interest

The main questions of interest are:

- Which variables/factors are associated with mental health outcomes such as life satisfaction, the WEMWBS score and the GHQ score?
- Are there any trends or changes in patterns of mental health as described by the above variables over the years of the Scottish Health Survey?

- Are there any differences in the above mental health outcomes between Glasgow and the rest of Scotland, after adjusting for demographic, socioeconomic and lifestyle factors?
- What other variables might be useful in better understanding what influences mental health in Scotland?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised Linear Models.
- Regression Modelling.
- Flexible Regression.

32 Self-rated health and socioeconomic status in Scotland (1)

Project level: Moderate

32.1 Overall project description

One of the questions in the Scottish Health Survey asks respondents to rate their own health. This assessment, known as self-rated health, can be very useful as it has been found to be strongly related to later illness and mortality.

Several factors are thought to influence poor health: broadly speaking, some of them are connected to social disadvantage, either present or in childhood; others are linked to lifestyle choices (such as diet, alcohol consumption, smoking habit, physical activity) which in turn may be affected by social circumstances.

This project will use data from the Scottish Health Survey of 2013 to investigate the association between self-rated health and social economic status (both present and in childhood) of the respondents, accounting for the possible influence of other behavioural factors.

32.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data from the 2013 Scottish Health Survey will be used in this project. For more information on the data and on how the variables are coded, please follow this link. The data are stored in the file `shs2013.Rdata` which contains the data frame `shs` with the following variables:

- `Sex` - Men (1), Female (2)
- `age` - Age of individual
- `totinc` - Total household income (ordered category, taking values 1-31, with 96 corresponding to 'Don''t know' and 97 to 'Refused'.)
- `pnssec5` - Parental National Statistics Socio-Economic Classification (NS-SEC) (highest) 5 groups
- `manssec5` - Mother's NS-SEC 5 groups (see `hpnssec5`)
- `fanssec5` - Father's NS-SEC 5 groups (see `hpnssec5`)
- `hedqul08` - Highest educational qualification (1 "Degree or higher", 2 "HNC/D or equiv", 3 "Higher grade or equiv", 4 "Standard grade or equiv", 5 "Other school level" 6 "No qualifications")
- `health` - Self-assessed general health, 1 for good, 0 for bad/fair
- `limitill` - Limiting long-standing illness (1 'Limiting LI', 2 'Non limiting LI', 3 'No LI')

- `hpnssec5` - Household representative person's (hrp) NS-SEC 5 variable classification (1 "Managerial and professional occupations", 2 "Intermediate occupations", 3 "Small employers and own account workers", 4 "Lower supervisory and technical occupations", 5 "Semi-routine occupations", 99 "Other".)
- `SIMD15_12` - Flag for Scottish Index of Multiple Deprivation 15% most deprived datazones
- `qsimd12` - Scottish Index of Multiple Deprivation quintiles, from 1 (least deprived) to 5 (most deprived)
- `drating` - Total Units of alcohol/week
- `drkcat3` - Weekly drinking category - 3 categories (1=non/2=moderate/3=hazardous or harmful)
- `cigst3` - Cigarette smoking status - 3 categories 1 "Current cigarette smoker", 2 "Ex-smoker", 3"Never smoked"

Question(s) of interest
The main questions of interest are:

- What is the effect of childhood socioeconomic status on health in adulthood?
- What is the effect of adulthood socioeconomic status on current health?
- Are childhood and adulthood socioeconomic effects independent?
- How do other variables relate to self-rated health?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised Linear Models.
- Regression Modelling.
- Flexible Regression.

33 Using the urinary steroid profile to detect prostate cancer in men (1)

Project level: Moderate

33.1 Overall project description

This project will focus on modelling the urinary levels of Endogenous Anabolic Androgenic Steroids (EAAS) for clinical purposes. The main aim is to explore whether EAAS could be used as a screening test for identifying metabolic imbalance and pathological conditions such as benign prostatic hyperplasia (BPH) and prostatic carcinoma. This would be an improvement over current diagnostic methods which are both more invasive and more expensive, and which do not perform particularly well in terms of diagnostic accuracy.

33.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The file `steroidprofile.csv` contains data on 518 men, who are either healthy (H), have benign prostate hyperplasia (BPH) or prostate cancer (CAP). Also available in the dataset are biomarker measurements taken from urine samples, such as Testosterone (T), Epitestosterone (E), Androsterone (A), Etiocholanolone (Etio), 5α-Androstane-3α, 17β-diol (5α Adiol), 5β-Androstane-3α, 17β-diol (5β Adiol), Dehydroepiandrosterone (DHEA), Dihydrotestosterone (DHT) and others. For more information on the biomarkers, see the first reference below. In addition, ratios such as T/E, A/T, A/Etio, 5α Adiol/5β Adiol and 5α Adiol/E are provided. For some of the subjects, we also have information on the subject's age.

The first observation is shown below.

```r
st <- read.csv("steroidprofile.csv")
head(st,1)
```

```
##   ID CLASS age X16a.OH.ANDROSTENDIONE X4.ANDROSTENDIONE X4.OH.TESTOSTERONE
## 1  1     H  16                   0.75         0.5741686           2.173123
##   X5aADIOL.5bADIOL X5aADIOL.E X5a.ADIOLO X5.ANDROSTENDIOLO X5b.ADIOLO
## 1         2.269738   1.371527   42.21045          13.80199   18.59706
##   X4.6.ANDROSTADIENDIONE X7a.OH.TESTOSTERONE X7b.OH.DHEA       A   A.ETIO
## 1              0.8644691           0.2945897    8.428744 824.716 1.626751
##        A.T DELTA6.TESTO     DHEA      DHT        E     ETIO FORMESTANO
## 1 832.6454          0.2 27.72504 8.558769 30.77624 506.9713   3.571328
##           T        T.E X6.DEIDROANDROSTERONE
## 1 0.9904769 0.03218317                    NA
```

Question(s) of interest

The main questions of interest are:

- For the healthy subjects with age information available, is there a relationship between age and the steroid profile?
- How well can the participants of this study be classified into Healthy, BPH patient or prostate cancer patient based on their urinary steroid profile?
- What other information is needed in order to compare your results with current diagnostic tests for prostate cancer?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.
- Generalised Linear Models.

34 Does playing Pokemon Go increase physical activity? (1)

Project level: Easy/Moderate

34.1 Overall project description

This project will analyse data from a published study on Pokemon Go players' attitudes towards exercise, playing frequency and physical activity.

34.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The file `pokemon.Rdata` contains the data frame `pok` with data on 999 Pokemon Go players from the US.

The first observation is shown below.

```
load("pokemon.Rdata")
head(pok,1)
```

```
##   id  submitdate          ipaddr age                      education Gender
## 1  7 12534912000 166.67.66.242    38 Some college credit, no degree Female
##   attitude_attitude1 attitude_attitude2 attitude_attitude3 attitude_attitude4
## 1     Strongly agree     Strongly agree           Disagree     Strongly agree
##   ATTENTION_filter1 attitude_attitude5 attitude_attitude6
## 1          Disagree     Strongly agree     Strongly agree
##   stepsattitude_attitudeB1 stepsattitude_attitudeB2 stepsattitude_attitudeB3
## 1                        7                        1                        7
##   stepsattitude_attitudeB4 stepsattitude_attitudeB5 stepsattitude_attitudeB6
## 1                        5                        3                        7
##   RecencypastBehavior_recencybike RecencypastBehavior_recencywalk
## 1         More than one month ago                       Yesterday
##   RecencypastBehavior_recencyrun perceivedBehav_freqWalking
## 1           During the last week          from 6 to 8 times
##   perceivedBehav_freqRunning perceivedBehav_freqBikeing
## 1                    2 times                      Never
##   app_usage_PokemonGoApp_pokemonusage1 social_sharing
## 1                            Sometimes   Occasionally
##   PokemonPastBehavior_pokPast1 PokemonPastBehavior_pokPast2
## 1            from 3 to 5 times                        Never
##   PokemonPastBehavior_pokPast3
## 1                        Never
##   PokemonPastBehavior_pokPast4_pokemonusage_NOT_USED
## 1                                          Sometimes
```

A detailed description of the data is given in the first reference below, and an analysis of the data is presented in the second reference.

Reading material (links)

Gabbiadini, A., Sagioglou, C., Greitemeyer, T. Original dataset used in the article "Does Pokemon Go lead to a more physically active life style?". Data in Brief, 20 (2018), 732-734

Gabbiadini, A., Sagioglou, C., Greitemeyer, T. Does Pokemon Go lead to a more physically active life style? Computers in Human Behavior, 84 (2018), 258-263.

Kaczmarek, L.D., Misiak, M., Behnke, M., Dziekan, M., Guzik, P. The Pikachu effect: Social and health gaming motivations lead to greater benefits of Pokemon GO use, Computers in Human Behavior, 75 (2017), 356-363.

Question(s) of interest

The main questions of interest are:

- Is a higher frequency of playing Pokemon Go associated with a higher amount of physical activity?
- Are Pokemon Go players more likely to participate in physical activity in general, or just in app-related activity?
- Are there are any variables that are associated with the amount of physical activity reported?
- Are attitudes towards physical activity associated with participants' gender or educational level?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Modelling.
- Generalised Linear Models.

35 Modelling wind speed at a windfarm site (1)

Project level: Moderate

35.1 Overall project description

Significant wind energy generation potential exists in regions where the mean wind speed is large. Of these regions, forecasts are only useful where there is significant inter-annual variability in wind speed. However useful seasonal forecasts of wind generation potential can only be made where there is both large mean and variability, and skill in wind prediction.

Direct forecasts of expected power generation are useful to the industry. However, there may be more sophisticated metrics more relevant to user decisions. One such metric is the percentage of time expected to be out of action due to wind speeds above kick-out speed (25ms-1). This is considered in the analysis below. There may be other metrics such as extreme lows/highs (droughts/floods) in power production (intensity and duration).

However, modelling wind speed is a challenging task because wind speed (and also wind direction) is highly intermittent, and wildly unpredictable. Though fairly accurate models for wind speed exists, they conventionally require data collected over a ten-year period), to appropriately capture seasonal yearly patterns, and overall long-term trends and patterns, to accurately predict wind speed behaviour over the expected life span of a wind farm, which is around 25 years

35.2 Individual project details

How many individual projects are available in this area: 1.

Data available
The wind speed (from now on referred to as WS) measurements consist of hourly observations obtained during a period of three consecutive years beginning the first day of 2011 and ending the last day of 2013. MERRA is a database constructed by NASA's Global Modelling and Assimilation Office; the purpose of the data is to make satellite data available to the wider scientific community working on climate research. For the purpose of this case study, only seven variables from the set will be considered: Humidity (MERRA.U), Meridional Velocity (MERRA.V), Relative Humidity (MERRA.RH), Pressure (MERRA.P), Temperature (MERRA.T), Wind Speed (MERRA.WS), and Wind Direction (MERRA.WD).

There are 9 columns in the dataset

- `WS` - On site wind speed in meters per second (m/s)
- `MERRA.U` - MERRA zonal velocity in meters per second (m/s)
- `MERRA.V` - MERRA meridional velocity in meters per second (m/s)
- `MERRA.RH` - MERRA relative humidity - Ratio (%)

- `MERRA.P` - MERRA pressure in Pascal (hPa)
- `MERRA.T` - MERRA temperature in Kelvin (K)
- `MERRA.WS` - MERRA wind speed in meters per second (m/s)
- `MERRA.WD` - MERRA wind direction (Degree _)
- `Date.Time` - Date and Time of observation (Year, Month, Day, Hour)

Question(s) of interest
The main questions of interest are:

- To examine the distribution of measured wind speed, and to asses the time series for seasonal patterns and any trend. Later analysis will not simply focus on the mean of the WS distribution but also the quantiles.
- To examine how well the MERRA modelled data and the observed wind speed agree.

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Time series.
- Flexible regression (quantile regression and extremes).

36 Exploring a household's energy consumption (1)

Project level: Moderate

36.1 Overall project description

Since moving house in 2008, a member of the School has been keeping a record of the readings of his electricity and gas meters, roughly monthly. In this project you are invited to explore these time series individually and/or together to come up with plausible explanations of the patterns that they display that you can justify statistically. These stories might relate to his individual household or the interaction of that household with the environment. Two interventions that might be of interest are the installation of loft insulation on 12th December 2012 and the installation of double glazing in the week of the 27th March 2017. Do they have any impact?

36.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The basic data consist of a table of three columns:

- `date`: The date of the reading.
- `gas`: The reading on the gas meter. (This is in awkward units: to see how to deal with them, see **Helpful Starting Hints** below. Also after the meter reaches 9999 it wraps round back to 0.)
- `electricity`: The reading on the electricity meter (in kWh).

They are stored in a comma-separated values file at
http://www.stats.gla.ac.uk/~vincent/STATS5029P/energy.csv.

Questions of interest

Possible questions of interest include:

- Are there regular patterns of energy usage in the two sources of energy separately?
- How can these patterns be modelled?
- In what ways are electricity and gas usage similar and different?
- Is there an effect of the introduction of loft insulation and/or double glazing?

Relevant courses

We recommend that you have taken the following courses to undertake this project:

- Regression modelling.

- Time series.

37 How well can you establish the geographical origin of a DNA sequence? (1)

Project level: Moderate

37.1 Overall project description

One morning, at a large international statistics conference, a body is found slumped over the lectern. From the murder scene, a sample of blood, which does not match the victim and hence is presumed to be from the perpetrator, is recovered. Mitochondrial DNA (mtDNA) is successfully extracted from the blood sample. The question to be investigated is: can any inference be made about where in the world the perpetrator came from?
DNA sequences differ between individuals and the different sequences occur at different frequencies in different populations. Databases of samples of sequences from around the world are available. If the perpetrator's sequence is common only in a restricted part of the world, the legal system could be on to a winner. For example, it could potentially be useful to the police in refining their pool of suspects.

DNA sequences can be thought of as a sort of high-dimensional multivariate data. In this project, you will investigate how well short mitochondrial DNA sequences allow the assignment of sequences to their population of origin.

In principle, you can use any classification approach that you think appropriate. The data has many variables so dimension-reduction techniques (such as principal components analysis, PCA) might be applied at the outset.

The first task will be to investigate whether a broad continental assignment is possible.

37.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data consist of short mitochondrial DNA sequences from 1394 subjects from different human populations. You are not given the raw sequences (strings of the letters A, G, C, T representing the chemical constituents, called nucleotides, of DNA: adenine, guanine, cytosine and thymine). Rather, for every position in the sequence in the sequence where there is variability between individuals in the sample, you are given information on which individuals share the same letter, as described below. The data table has 1394 rows (subjects) and 206 columns (variables) as follows.

- Column `Continent`: Specifies the broad continent of the subject (AFR = Sub-Saharan Africa, ASI = East Asia, EUR = Europe)

- Column `Population`: Specifies a narrower population label (MAN = Mandenka [Senegal], MOZ = Mozambique, WAT = East Africa; CHI = China; JAP = Japan; BUL = Bulgaria, COR = Cornwall [UK], CZE = Czech Republic, FRA = France, WAL = Wales [UK], ITA = Italy).
- Columns 3 to 206: Each column represents a variable position in the DNA sequences. (The column title identifies that position, but it probably is of no interest.) A zero represents one nucleotide; a one represents a different nucleotide (it does not matter which).

They are stored in a comma-separated values file at
http://www.stats.gla.ac.uk/~vincent/STATS5029P/mtdna.csv.

The bibliographic details of the data are listed in the following:
http://www.stats.gla.ac.uk/~vincent/STATS5029P/mtdnasources.pdf.

Questions of interest
Possible questions of interest include:

- Do there appear to be systematic genetic difference between the three continental groups (after reducing the number of variables)?
- How well in general can individual sequences be assigned to continents?
- Is there an optimal degree of dimension reduction that makes classification as good as possible?
- Do there appear to be systematic genetic difference between the 11 population groups (after reducing the number of variables)
- How well in general can individual sequences be assigned to populations?

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Multivariate methods.
- Statistical Genetics is not required.

38 Supervised statistical classification (2)

Project level: Moderate

38.1 Overall project description

An important problem in data science is supervised classification, where the aim is to assign labels to instances described by a vector of feature variables. The classification task is guided by a statistical model learnt using data containing labelled instances. The array of models now designed to perform classification is constantly increasing. In this project you will compare the classification performance of some standard methods with more advanced ones of your choice. Some references that could be useful for the project are:

- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24;
- Pérez-Ortiz, M., Jiménez-Fernández, S., Gutiérrez, P., Alexandre, E., Hervás-Martínez, C., & Salcedo-Sanz, S. (2016). A review of classification problems and algorithms in renewable energy applications. Energies, 9(8), 607;

but many other books and/or articles can be consulted for an introduction.

38.2 Individual project details

How many individual projects are available in this area: 2.

38.3 Binary classification

Data available
The file Binary_Classification.zip includes four datasets from the UCI Repository https://archive.ics.uci.edu/ml/index.php, namely *Breast Cancer Wisconsin (Original)*, *Mammographic Mass*, *Tic-Tac-Toe Endgame* and *Wilt*. All these datasets have a binary class variable and a set of features. For both the *Breast Cancer Wisconsin (Original)* and the *Mammographic Mass* datasets the aim is to classify breast tumours as either benign or malign given a set of characteristics; for the *Tic-Tac-Toe Endgame* dataset the aim is to predict whether the player has won or not given the board configurations; for the *Wilt* dataset the aim is to detect diseased trees given some information from image segments. Additional details about these can be found in the UCI Repository. Although some datasets are provided, many others from the UCI repository could be equally used (as long as the class variable is binary).

Question(s) of interest
The main questions of interest are:

- What is the classification capability of the simple logistic regression model?

- Do other statistical models provide a better classification performance than logistic regression?
- How can a study can be designed to assess performance in classification? What measures of performance can be used?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Multivariate methods.

38.4 Multinomial classification

Data available
The file Multinomial_Classification.zip includes four datasets from the UCI Repository https://archive.ics.uci.edu/ml/index.php, namely *Abalone*, *Car Evaluation*, *Contraceptive Method Choice* and *Nursery*. For the *Abalone* dataset the aim is to predict the age of abalone from physical measurements; for the *Car Evaluation* dataset the aim is to predict the car acceptability given its features; for the *Contraceptive Method Choice* the aim is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics; for the *Nursery* dataset the aim is to predict whether applications to nursery school were successfull given socio-demographic information of the parents. All these datasets have a multinomial class variable and a set of features. Details about these can be found in the UCI Repository. Although some datasets are provided, many others from the UCI repository could be equally used (as long as the class variable has more than two levels).

Question(s) of interest
The main questions of interest are:

- What is the classification capability of the simple multinomial regression model?
- Do other statistical models provide a better classification performance than multinomial regression?
- How can a study can be designed to assess performance in classification? What measures of performance can be used?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Multivariate methods.

39 Spatio-temporal modelling of big environmental data (3)

Project level: Moderate

39.1 Overall project description

Data of environmental interest, such as river flows or daily average temperatures, are usually collected at multiple stations over a geographic region of interest and over time. Interest is then in understanding how the response of interest varies over both space and time as well as over other covariates that may be relevant. Data of this type abound nowadays and has been collected over long periods of time. In this project you will apply appropriate statistical methods to long time-series of data coming from real environmental applications.

39.2 Individual project details

How many individual projects are available in this area: 3.

39.3 Average daily temperatures in the US

Data available

The file `ustemp.csv` includes the average daily temperatures at 48 US cities between 01/01/1995 and 31/12/2018, for a total of 420768 observations together with covariates that may affect the response. The data is publicly available from the Average Daily Temperature Archive of the University of Dayton. A description of the data can be found at http://academic.udayton.edu/kissock/http/Weather/default.htm. The dataset `ustemp.csv` includes the variables:

- `City`: name of the city;
- `Temp`: average daily temperature;
- `Date`: date of the temperature recording;
- `Lat`: latitude of the city;
- `Long`: longitude of the city;
- `Alt`: altitude of the city;
- `Sea`: whether the city is by the coast (coded as 1) or not (coded as 0);

but others could be added if needed.

Question(s) of interest

The main questions of interest are:

- Do average temperatures in the US vary over space?
- Do average temperatures in the US vary over time?
- Are the covariates effective to predict average temperatures?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.
- Time series.
- Environmental statistics.
- Flexible regression.

39.4 Average daily river flows in Scotland

Data available

The file `River_flows.csv` includes the average daily flows of 64 Scottish rivers between 01/01/1989 and 31/12/2015, for a total of 640575 observations together with covariates that may affect the response. The data is publicly available from the National River Flow Archive. A description of the data can be found at https://nrfa.ceh.ac.uk. The dataset `ustemp.csv` includes the variables:

- `ID`:ID of the station;
- `Date`: date of the flow recording;
- `Flow`: average daily flow;
- `Station`: name of the river;
- `Latitude`: latitude of the station;
- `Longitude`: longitude of the station;
- `Easting`: easting of the station;
- `Westing`: westing of the station;
- `Catchment.Area`: catchment area of the measured river;
- `Max.Altitude`: max altitude of the measured river;

but others could be added if needed. Information about the data can also be found in

- Franco-Villoria, Maria, Marian Scott, and Trevor Hoey. Spatiotemporal modeling of hydrological return levels: A quantile regression approach. Environmetrics 30.2 (2019): e2522.

Question(s) of interest

The main questions of interest are:

- Do average daily flows in Scotland vary over space?
- Do average daily flows in Scotland vary over time?
- Are the covariates effective to predict average flows?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.
- Time series.
- Environmental statistics.
- Flexible regression.

39.5 Maxima daily temperatures in under canopy vs. open field stations in Switzerland

Data available

The file `Swiss.csv` includes the maxima daily temperatures at 28 recording stations over 14 sites in Switzerland between 01/01/2002 and 31/12/2015, for a total of 142828 observations together with covariates that may affect the response. Each site consists of two stations, one in open-field and the other under the forest canopy. The data is publicly available from the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL). A description of the data can be found at https://www.wsl.ch/en/forest/forest-development-and-monitoring/long-term-forest-ecosystem-research-lwf.html. The dataset `Swiss.csv` includes the variables:

- `station`: name of the site;
- `date`: date of the temperature recording;
- `temp`: recorded maxima daily temperature;
- `latitude`latitude of the site;
- `longitude`: longitude of the site;
- `altitude`: altitude of the site;
- `slope`: slope of the site;
- `type`: type of the station (field or forest)

but others could be added if needed. Information about the data can also be found in

- Renaud, V., et al. Comparison between open-site and below-canopy climatic conditions in Switzerland for different types of forests over 10 years (1998 - 2007). Theoretical and Applied Climatology 105.1-2 (2011): 119-127.

Question(s) of interest

The main questions of interest are:

- Do maximum temperatures in Switzerland vary over space?
- Do average temperatures in Switzerland vary over time?
- Are the covariates effective to predict average temperatures?
- Is there a difference in temperatures between open-field and under forest canopy stations?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Data Analysis.
- Time Series.
- Environmental statistics.
- Flexible Regression.

40 Modelling sunspot numbers (1)

Project level: Moderate/Difficult

40.1 Overall project description

Our sun in a volatile system. Yet it displays at least one striking pattern, a quasi-periodic variation in its magnetic activity, called the solar cycle. One visible manifestation of this is in the number of sunspots (dark patches on the surface, which are quite easy to observe, given suitable precautions): observations extend back more than 300 years. This number oscillates with a period of roughly 11 years (i.e., with a frequency of roughly 1/11 cycles per year). We are currently in a time of increasing numbers of sunspots.

In this project you are invited to model this periodic behaviour, explore its relationship to the solar magnetic field and also to explore whether you can find any influence of the solar cycle on the earth, for example on the temperature.

40.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The sunspot data are available from the **World Data Center for the production, preservation and dissemination of the international sunspot number** (SILSO) at http://sidc.be/silso/datafiles. They are reported daily, but you might wish to work with yearly or monthly mean data. Daily data go back to 1818; monthly means are available back to 1749 and yearly means go back to 1700.

Questions of interest

Possible questions of interest include:

- Can you estimate the apparent periodicity in the sunspot number more precisely than just "roughly 11 years' '?
- What is the relationship between sunspot number and solar magnetic field?
- Does the observation that the solar magnetic field flips in sign every other cycle of the sunspot numbers suggest a better model of the sunspot numbers?
- Does the solar cycle impact average temperature on earth?

Relevant courses

We recommend that you have taken the following courses to undertake this project:

- Regression modelling.
- Statistical Inference.

- Time series.

41 Using simulation to investigate properties of the likelihood ratio (or G) test for Hardy–Weinberg equilibrium in genetics (1)

Project level: Moderate

41.1 Overall project description

Hypothesis tests use the distribution of a test statistic under a null hypothesis to establish whether it is plausible that the observed value of that test statistic (that comes from a real data set) has been generated from that null hypothesis. If it isn't plausible, the null hypothesis is rejected. In very simple cases (e.g., the one-sample t test), there are exact formulae for that distribution. In other cases there are approximate formulae. In others, no formulae at all! This project looks at a particular hypothesis test (in the context of genetics) for which approximate results are available and asks you to see how good the approximation is.

Under strong assumptions, the genotype frequencies of a gene reach an equilibrium named after Hardy and Weinberg (HW). If the organism has two copies of the gene (called diploid, like humans) and if there are two different forms of the gene, the alleles A and a, the genotype relative frequencies of AA, Aa, aa when equilibrium has been reached should be p^2, $2pq$, q^2, respectively, where p is the relative frequency of the allele A and $q = 1-p$. Observed genotype frequencies are routinely compared to these HW proportions and a G test (a likelihood ratio test for multinomial data) is typically performed. This relies on the result that, for large enough sample sizes, the G statistic is approximately chi-squared distributed.

This project investigates whether that chi-squared distribution assumption is good enough in practice, using simulation. This is important because so many of these tests are done given the large number of genes that are typically assayed in current genomic studies.

One assumption that is needed for the population to reach Hardy-Weinberg equilibrium (HWE) is that the population is randomly mating. That is, individuals choose their mates at random from the population. If this is not true and there are really subpopulations from which one is more likely to choose one's mate, it can be shown that this decreases the number of observed heterozygotes (those with genotype Aa).

The second part of this project investigates the power of the G test to detect this decrease in heterozygotes, again by simulation.

41.2 Individual project details

How many individual projects are available in this area: 1.

Data available
This is a simulation-based project.

Questions of interest
Possible questions of interest include:

- Under what conditions is the chi-squared distribution assumption for G poor?
- Does using it cause the test to make it more or less likely to reject the null hypothesis?
- If population structure is present, what is the power of the G test to detect it, for varying levels of structure?

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Statistical Inference.
- Statistical Genetics.

42 Exploring variation in human skull shape (2)

Project level: Moderate

42.1 Overall project description

Physical anthropologists spend a lot of time measuring human bones (ancient and modern) to learn about how variation in anatomy is distributed across populations. Such painstaking work has provided an important source of stories about human evolution and dispersal.

These projects will explore a large data set of human cranial measurements that consists of samples from diverse human populations with many measurements on each skull. The measurements are typically distances between well-defined landmarks on the skull.

The projects will investigate how population and sex affect the measurements and how the measurements can (or cannot) be used to classify subjects by sex or population.

42.2 Individual project details

How many individual projects are available in this area: 2.

Data available
The data consist of measurements between pairs of landmarks (well-defined features that experts can locate on a skull) on the crania of 2524 humans from 28 populations.

The data can be downloaded here: https://web.utk.edu/~auerbach/HOWL.htm. It is the **Howells Craniometric Data Set**.

- The first (ID) column is a sample code.
- The sex is in the second column.
- Each population has a number and a name in the third and fourth columns, respectively.
- The remaining 82 columns contain different measurements on the skulls (each with a name).

Important note: a measurement recorded as zero is a missing measurement, not a genuine value 0, so it is an NA in R-speak.

42.3 Regression/Analysis-of-Variance approaches to the cranial data

Questions of interest
Possible questions of interest include:

- Is there an effect of sex on the different measurements?

- Is there an effect of population on the different measurements, allowing for sex?
- Do any population effects on the measurements depend on sex?
- Can you build a parsimonious model to predict sex from cranial measurements?

42.4 Classification approaches to the cranial data

Questions of interest

Possible questions of interest include:

- How well can individual samples be assigned to their sex?
- How well can individual samples be assigned to their population?
- How well can individual samples be assigned to larger "superpopulations"?
- Is it helpful to perform dimension reduction before doing classification?

Relevant courses

We recommend that you have taken the following courses to undertake these projects.

- Statistical Inference.
- Regression Modelling.
- Generalised Linear Models.
- Multivariate methods.

43 Usage of hire bikes in London (1)

Project level: Moderate/Difficult

Target degree programme: Likely most suitable for MSc in Data Analytics

43.1 Overall project description

As part of its open data framework Transport for London (TfL) released anonymised trip data for the public cycle hire scheme. TfL provide for every trip the location and time the bike was taken out as well as returned (data is available at http://cycling.data.tfl.gov.uk/).

In this project you will visualise and analyse this data set and look for trends and patterns.

A key challenge of this project is that it involves big data (a week's worth of data is around 20MB). There is a wealth of information available, which needs to be distilled down to a smaller amount of information relevant to answering the questions of interest.

43.2 Individual project details

How many individual projects are available in this area: 1.

Data available
Initially, the number of trips for each hour of each day of the year 2016 is available as a `.RData` file, called `all_trips.Rdata`. This file contains the following 3 columns in the `R` object `data`:

- Date: the calendar day of the year;
- Hour: the hour of the day;
- Trips: the number of trips within a particular hour on a particular day.

The full data for this project can be downloaded from the TfL website. Specifically, the full Cycle usage data can be downloaded from https://cycling.data.tfl.gov.uk/. The usage data is available towards the bottom of the page.

Question(s) of interest
For the available 2016 data, initially:

What are the detailed spatio-temporal patterns? What time of the day is most popular? Are weekdays different from weekends? Is summer different from winter?

The project can then be extended to look at the above questions of interest for all data and additionally to explore: Which bike stations are the most popular? What types of trips are the bikes used for?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Generalised Linear Models.
- Time Series.
- Flexible Regression.

44 Local elections in England – What is the message? (1)

Project level: Moderate

44.1 Overall project description

In recent local elections in England in 2019, the Conservatives fared very badly, but Labour did not do all that well either. On the other hand, the Liberal Democrats, the Greens and independents did very well.

Politicians wonder, and argue about, what the message from voters was. Was the message to the Tories to press ahead with a hard Brexit? Were former Labour votes disappointed by Labour's (lack of) stance on Brexit? Do they want Labour to become more forceful proponents of a second referendum?

We cannot fully answer those questions using data, but we can try to relate the changes in voting patterns to the results of the EU referendum in that local authority as well as other administrative data, such as the distribution of social grades.

44.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data available consist of

- a data frame containing the proportion of the population for each social grade (AB, C1, C2, DE), the results from the last general election (Westminster) as well as the result from the EU referendum (proportion leave votes).
- a list containing the results from the local election scraped from the BBC website a few days after the election.

Additional administrative data is available from the Office of National Statistics.

Question(s) of interest

The main question of interest is:

What are the characteristics of local authorities …

- where the Conservatives lost a large proportion of the seats up for election?
- where Labour lost a large proportion of the seats up for election?
- in which the Liberal Democrats / Greens / independents made the largest gains?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Generalised Linear Models.
- Flexible Regression.

45 Predicting astrophysical properties of stars based on their light spectrum (1)

Project level: Moderate

45.1 Overall project description

Gaia is an ESA space mission launched in 2013. Its objective is to compile a catalogue of approximately 1 billion stars, roughly 1% of the stars in the Milky Way. The satellite will be equipped with spectrophotometric detectors (essentially sophisticated versions of the CCD chips found in digital cameras). The light spectra of each star can be used to predict astrophysical properties such as the effective temperature, the surface gravity and the stellar metallicity (logarithm ratio Fe/H).

The relationship between these properties and the light spectra has to be learned from simulated data. The actual data measured by the satellite does not contain the temperature, surface gravity and stellar metallicity and thus cannot be used for learning or assessing the quality of the models. For this reason the project will only work with data from a complex astrophysical simulation model (photosim/BaSeL 2.2).

A description of the data is available at http://www2.mpia-hd.mpg.de/Gaia/icap/Simulated_data.shtml.

45.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data contains (amongst other) the three columns we want to predict (`temperature`, `gravity` and `metallicity`, as well as 16 columns of normalised photon counts for different wavelength buckets (called `count1` to `count16`).

Question(s) of interest

The main question of interest is to predict the three properties. The main focus should be predictive performance rather than interpreting the models fitted. An honest quantification of the uncertainty of the predictions would also of great use in the context of this project.

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Multivariate Methods.
- Machine Learning.

46 Housing market analysis (3)

Project level: Easy

46.1 Overall project description

House prices and predictors of house prices are well studied economic indicators. This dataset contains several variables which are related to house price and are collected at the time of house sale by a firm. The housing data ([`housing.csv`]) collected by the firm includes 500 sales in the last six months and include the following variables.

- **elevation**: Elevation of the base of the house
- **dist_am1**: Distance to Amenity 1
- **dist_am2**: Distance to Amenity 2
- **dist_am3**: Distance to Amenity 3
- **bath**: Number of bathrooms
- **sqft**: Square footage of the house
- **parking**: Parking type
- **precip**: Amount of precipitation
- **price**: Final House Sale Price
- **asking**: Indicator of whether the house sold above asking price or below asking price.

Most of these data are collected from real estate databases and indvidual buyers, sellers and real estate agent might use different features of this data for their planning

46.2 Individual project details

How many individual projects are available in this area: 3.

46.3 Best possible regression

Data available

Data on 500 house sales are available in the `housing.csv` file. In this project we are primarily interested in predicting the house sale price from other explanatory variables. You are allowed to choose between any regression model and use all other variables as possible predictor(s)

Question(s) of interest

The main questions of interest are:

- Are there any obvious outliers and how do you deal with them?

- What is the effect of each of the different predictors on the Final House Sale Price?
- Do you need to transform any variables? If so explain why?
- What is the best model for predicting the Final House Sale Price?
- Is the best model a linear model or a more general model?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Modelling.
- Generalised Linear Models.
- Advanced Data Analysis.
- Big data Analysis.

46.4 Classification

Data available
Data on 500 house sales including the variable whether the house sold below or above the asking price are available in the \texttt{housing_new.csv} file. In this project we are primarily interested in classifying which houses sold above asking price and which ones sold below and the response variable is `asking` price.

Question(s) of interest
The main questions of interest are:

- What model is most appropriate for modeling this binary random variable?
- What variables do we need to model this binary random variable?
- Can one use other classfication techniques such as Random Forest, Classification trees, Neural networks to provide a classification tool to model which houses will sell above the asking price
- To evaluate your methods provide a detailed comparison based on using the first 400 data as training and the last 100 as test.

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Modelling.
- Generalised linear models.
- Advanced Data Analysis.

46.5 Clustering

Data available
This project relates to finding clusters in the housing dataset. The housing data has several variables and the goal of this project is to find natural clustering in this market. Real estate agents and buyers are often interested in these clusters and might focus on one of these clusters for their house search.

Question(s) of interest

The main questions of interest are:

- Are there any natural clusters in the real-estate market?
- Do we need all variables to provide clustering?
- What type of clustering methods provide the best interpretation of the data?
- How do you compare among clustering methods?
- How do you choose the number of clusters?
- How do you interpret the clusters?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Analysis.
- Advanced Data Analysis.

47 Grocery sales data analysis (3)

Project level: Easy

47.1 Overall project description

Understanding grocery sales data is very important. The variables are

- **Weight** : Weight of product
- **Type** : The category to which the product belongs
- **Price** : Maximum Retail Price (list price) of the product
- **Promotion**: whether promotion was running on the product (1- yes, 0 - No)
- **Location** : The type of city in which the store is located
- **Outlet** : Whether the outlet is just a grocery store or some sort of supermarket
- **Sales** : Sales of the product in the particular store.

47.2 Individual project details

How many individual projects are available in this area: 3.

47.3 Regression Analysis

Data available

Data on 7060 product sales are available in the `product.csv` file. In this project we are primarily interested in predicting the total sale of the product (variable `Sales`) in the store from other explanatory variables. You are allowed to choose between any regression model and use all other variables as possible predictor(s)

Question(s) of interest

The main questions of interest are:

- Are there any obvious outliers and how do you deal with them?
- What is the effect of each of the different predictors on the Sales of the product in the particular store?
- Do you need to transform any variables? If so explain why?
- What is the best model for predicting the Sales of the product in the particular store?
- Is the best model a linear model or a more general model?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Modelling.

- Generalised Linear Models.
- Advanced Data Analysis.
- Big data Analysis.

47.4 Classification

Data available
Data on 7060 product sales are available in the `product.csv` file. In this project we are primarily interested in the outcome variable `promotion`. We are interested in finding out which factors contribiute to stores running promotion on items.

Question(s) of interest
The main questions of interest are:

- What model is most appropriate for modeling this binary random variable?
- What variables do we need to model this binary random variable?
- Can one use other classfication techniques such as Random Forest, Classification trees, Neural networks to provide a classification tool to model which houses will sell above the asking price
- To evaluate your methods provide a detailed comparison based on using the first 6000 data as training and the last 1010 as test.

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Modelling.
- Generalised linear models.
- Advanced Data Analysis.

47.5 Clustering

Data available
This project relates to finding clusters in the product sales dataset. The product sales dataset have several variables and the goal of this project is to find natural clustering based on the three continous variables , `Price, Sales` and `Weight`. Store owners are often intested in these clusters and might focus on stocking specific type of products. For the advance task you can include other variables.

Question(s) of interest
The main questions of interest are:

- Are there any natural clusters in the real-estate market?
- Do we need all variables to provide clustering?
- What type of clustering methods provide the best interpretation of the data?
- How do you compare among clustering methods?
- How do you choose the number of clusters?
- How do you interpret the clusters?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Analysis.
- Advanced Data Analysis.

48 Analysing trends in water quality in the Clyde estuary (1)

Project level: Moderate

48.1 Overall project description

The Scottish Environment Protection Agency (SEPA) has a statutory obligation to monitor the state of the environment. As part of that duty, water samples are taken regularly from sampling stations along the Clyde River. Data are available for a twenty year period from the mid-1970's until the mid-1990's. A natural measure of water quality is dissolved oxygen (DO). There is interest in identifying the pattern of DO along the river, the nature of any time trends and the relationship between DO and physical variables such as temperature and salinity. Data are also available at different depths. The aim of the project is to use this information to provide a description of how the health of the estuary has changed over this 20 year period.

48.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on dissolved oxygen at a variety of sampling stations up and down the river and at different times of year. Measurements of temperature and salinity are also available. Over the 20 year period there are 11085 observations. The data are stored in the text file `clyde.dat`. The code below reads this and creates a dataframe.

```
temp     <- scan("clyde.dat", na.strings = "*")
temp     <- as.numeric(temp)
d        <- matrix(temp, ncol = 15, byrow = T)
d        <- as.data.frame(d)
d        <- d[ , 1:13]
names(d) <- c("Station", "Day", "Month", "Year", "Stime", "HWGMT",
              "LWGMT", "Tidal", "Depth", "Temp", "Salinity", "DO", "sat")
d$id     <- factor(d$Day * 10000 + d$Month * 100 + (d$Year - 1900))
d$doy    <- cumsum(c(0, 31, 29, 31, 30, 31, 30, 31, 31, 30, 31, 30))[d$Month] + d$Day
d$year   <- d$Year + d$doy / 365
```

The dataframe `d` contains the following columns:

- `DO` - a measurement of dissolved oxygen.
- `Station` - the location of the sampling station, expressed as the number of miles downstream from the city centre.

- `Day, Month, Year` - the date the measurement was made.
- `doy` - the day of the year (0 - 365) the measurement was made.
- `year` - the time of the measurement on the year scale, including the proportion of time through the year corresponding to the day of the measurement.
- `Depth` - the water depth at which the measurement was made.
- `Temp` - the water temperature of the water sample.
- `Depth` - the salinity of the water sample.
- `id` - an identifier of the survey on which the water sample was taken.

The other variables may be ignored.

Question(s) of interest

The main questions of interest are:

- What are the main trends in water quality over this 20 year period?
- How do these trends differ at different sampling stations and at different times of year?
- What is the influence of temperature and salinity?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Regression models.

49 Predicting hotel booking cancellations (1)

Project level: Moderate

49.1 Overall project description

Our goal is to use a predictive model, such as a **decision tree**, to allow hotel managers to:

- accurately predict net demand and **build better forecasts**,
- **improve cancellation policies**, and/or
- **define better overbooking tactics**.

As a result, hotel management can use more assertive pricing and inventory allocation strategies.

49.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data set refers to hotel bookings from July 2015 to August 2017. Specifically:

- It refers to two Portuguese hotels (one in Algarve and one in Lisbon).
- It's comprised of 31 variables describing the $119,390$ observations.
- Data elements of hotel and/or customer identification are not available.

Question(s) of interest

The main questions of interest are:

- Can we predict hotel booking cancellations?
- If so, which variables are the ones that can be most helpful in predicting cancellations?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.

50 Multivariate disease mapping of three of Scotland's biggest killers (1)

Project level: Moderate

50.1 Overall project description

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 (CHD), Cerebrovascular Disease (CVD) 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.

50.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The dataset contains counts of the number of hospital admissions for coronary heart disease, cerebrovascular disease and respiratory disease in each intermediate zone (IZ) in Scotland in 2012. The country is divided into 1235 IZs which are non-overlapping small administrative areas, which contain on average 4000 household residents. The dataset also contains a number of covariates relating to each region.

This data was obtained from the Scottish government via https://statistics.gov.scot/home. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `MultiDisease.csv` and contains the following columns.

- `IZ` - The intermediate zone code.
- `Ei` - The expected number of hospital admissions for each IZ.
- `Yi` - The observed number of hospital admissions for each IZ.
- `JSA, Urban, Percent.Asian, Percent.Black` - Selected covariates for each zone - these include the percentage of 16-64's claiming job seekers allowance for each IZ, whether an area is deemed to be urban (1) or rural (2), the percentage of people of Asian ethnicity in each IZ and the percentage of people of Black ethnicity in each IZ.
- `Disease` - The disease which the hospital admissions relate to.

Question(s) of interest
The main questions of interest are:

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

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.
- Multivariate Methods.

51 Determining the drivers of yeast cell cycle gene regulation (1)

Project level: Moderate/Difficult

51.1 Overall project description

A cell cycle is a sequence of events by which a cell grows and divides into daughter cells that each contain the information and necessary apparatus to repeat the process. The most important component is the DNA, or genetic material present in chromosomes, which must be replicated accurately, with the two copies being carefully placed in the two new cells. The different processes involved- DNA replication, and separation into daughter chromosomes- occur in temporally distinct phases of the cell cycle: the S-phase (for "synthesis"), and M-phase (for "mitosis"), separated by two gaps, called G1 and G2. The information required to initiate and conduct these processes in encoded in the DNA, and this is decoded and used during the process of gene expression, which can be measured through laboratory experiments using a tool called a DNA microarray. Gene expression itself is controlled by regulatory proteins in the cell, called "transcription factors". Transcription factors regulate the expression of a gene by binding to a location in the gene regulatory region on the DNA sequence, and this location typically has a high degree of sequence specificity for any particular transcription factor. Determining which combinations of transcription factors are active during different parts of the yeast cell cycle is important in understanding not just the workings of the cell cycle itself, but could give insights into the general development of multicellular organisms.

51.2 Individual project details

How many individual projects are available in this area: 1.

Data available

There are two data sets for this project.

The first data set, called `YeastGeneExp.csv` comes from a gene expression study in yeast, with each row corresponding to a gene, and contains the following columns.

- Column 1: `ORF` - A unique name for each gene
- Columns 2-19: `alpha0`, `alpha7`, … `alpha119`
  - Gene expression measured for each gene at 18 time points in the cell cycle. (The data has been pre-processed and normalized to remove experiment-specific biases and measurement error, as far as possible.)
- Column 20: `phase`

  - Phase of the cell cycle at which the gene is known to be most active: options are S, M, G1, G2, and overlaps between neighbouring phases

The second data set, called `YeastTFscore.csv` details for each of the genes in the data set, a sequence match or binding "score" between the transcription factor and the regulatory region of the gene. A higher binding score often indicates a higher likelihood of the gene being regulated by the transcription factor, but this is not always true.

- Columns 1-32: `T1` … `T32`
- sequence match scores for each gene, for 32 transcription factor candidates.

Question(s) of interest
The main questions of interest are:

- How do gene expression patterns vary over the yeast cell cycle?
- Can the gene expression patterns be clustered correctly into their cell-cycle phase groups?
- In any particular group of cell-cycle genes, what are the transcription factors that are likely to be involved in their regulation?
- Which transcription factors affect the regulation of genes, and at which points, over the whole of the yeast cell cycle?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Time series.
- Regression modelling.
- Multivariate methods or Machine Learning.
- Any two of: Stochastic processes, Functional data analysis, Bayesian Statistics, Flexible Regression, Multivariate methods or Machine Learning.

52 Using Earth Observation to Identify Coherence in Surface Temperature of the Caspian Sea (1)

Project level: Moderate/Difficult

52.1 Overall project description

The quantity of data we are collecting is increasing at an unprecedented rate with the advent of new Earth Observation (EO) technologies that obtain data on our environment using satellites. These new data sets enable us to use statistical models to explore and describe changes in our natural environment.

One way to gain a better understanding of our environment is to explore coherence within environmental variables via unsupervised learning methods, such as clustering, in order to identify groups of individuals that share similar characteristics.

Specifically, this project aims to investigate and cluster features of lake surface water temperature (LSWT) within the largest land locked water body on Earth, the Caspian Sea. The physcial isolation of the Caspian Sea basin has created a unique ecological system which makes it of great interest for study. Temporal coherence can be defined as the synchrony between major fluctuations in a set of time series. In this project, the aim is to identify temporal coherence of LSWT at the Caspian Sea.

52.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data have been collected via Earth Observation and come from the ArcLake project (http://www.laketemp.net/home_ARCLake/) which has processed data collected from the European Space Agency's Adavanced Along Track Scanning Radiometers.

There are 1990 time series of LSWT available, each with 405 bi-monthly values covering the time period from June 1995 to April 2012. Each of the 1990 time series corresponds to a grid square location on the surface of the Caspian Sea. These grid squares are known as pixels and correspond to an area which is approximately $3km^2$.

The data are stored in `caspian.csv` which contains the following columns.

- `cellids` - A unique code for each location.
- `x` - the x co-ordinate (longitude)
- `y` - the y co-ordinate (latitude)
- `1995-06-08 ... 2012-04-08` - columns 4 to 408 contain the LSWT values on the date corresponding to the column names for each pixel. LSWT is measured in Kelvin.

i.e. The rows correspond to locations (pixels) and columns (from column 4-408) correspond to the time series.

Question(s) of interest
The main questions of interest are:

- What are the key patterns over time in LSWT in the Caspian Sea?
- Can we identify any clusters of common temperature patterns within these lakes? If yes, what is the optimal number of clusters within needed to describe the underlying temporal temperature dynamics.
- How are these clusters distributed spatially across the area of interest?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods (Essential).
- Flexible Regression (Desirable).
- Functional Data Analysis (Desirable).

53 Understanding trends in garden bird counts using zero-inflated models (1)

Project level: Moderate/Difficult

53.1 Overall project description

The provision of food for birds in gardens is a multimillion pound industry with great public investment in the numbers. It has also been shown that garden birds give a good indication of how the wider environment is performing. The British Trust for Ornithology has been running the Garden Bird Feeding Survey since 1970, where amateur scientists count the number of birds seen feeding on food they put out in the garden. As averaged counts, these are better treated as a continuous variable rather than discrete, but there are many more exact zeros than can be traditionally accounted for using standard continuous distributions. Careful choices of probability distribution must therefore be made to account for this (Tweedie (`fishMod` package) and delta lognormal (`EnvStats` package) models are potential candidates). This project will construct statistical models to analyse data on garden bird abundance with the aim of determining the important predictors of positive and/or negative changes in abundance.

53.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on numbers of annual averaged counts of robins and starlings, across 693 monitoring sites, spanning 45 winters. There is also a selection of environmental and spatial variables for the same location and year. The data are stored in `Garden_bird.csv` and contain the following columns.

- `site` - A unique code for each IZ area.
- `year` - The year of observation (year 1 being the winter of 1970/71)
- `subrur` - A two-level variable stating whether the site is (sub)urban (+1) or rural (-1).
- `easting` - A standardised longitude measure.
- `northing` - A standardised lattitude measure.
- `temp,frost` - Measures of average temperature/number of days of ground frost respectively over the winter months.
- `sparr` - Averaged number of sparrowhawks (a predator) observed, considered by some to be a potential important driver of population changes.
- `robin,starl` - The response variables (averaged counts of robins and starlings respectively).

Question(s) of interest

The main questions of interest are:

- What are the important variables in predicting changes in bird numbers?
- Which areas have the highest and lowest bird densities/greatest change in numbers?
- Are there differences in inference using different probability distributions?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Advanced Bayesian Methods.
- Spatial statistics.
- Time series.

54 Statistics Anxiety in Introductory Statistics Courses (1)

Project level: Easy

54.1 Overall project description

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 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 use 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 of three different introductory statistics courses: the first ("S1Y") intended for specialists in the mathematical sciences, the second ("Bio") a compulsory course for specialists in another field and the third ("S1A") an optional course for specialists in another field. In addition, the specialist course has comparable data on students' perceptions of statistics at the start and end of the course. Appropriate ethical approval and participant consent were obtained in all three courses.

54.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available for each course describing the students' demographics, self-efficacy, statistical anxiety and general perceptions. Data on the performance of each student in the each course is also included. The data are stored in *Statistics_Anxiety_Project.csv* and contain the following columns.

- `ID` Unique identifier for each student
- `Gender` i.e. "Female","Male", "Other"
- `Age` Age of student at the start of the course
- `Nationality` e.g. "British", "Spanish", "Cypriot", "American", …
- `College` i.e. "Arts","MVLS", "Science & Engineering", "Social Science"
- `Year.of.Study` i.e. "Year 1", "Year 2", "Year 3", "Year 4", "Masters year"

- `Subject` Main subject of study, e.g. “Maths”, “Mathematics and Statistics”, ...
- `Anxiety.Level` Response to “Please rate your levels of anxiety towards studying statistics?” where 1 = “Not Anxious”, 2 = “A little anxious”, 3 = “Quite anxious”, 4 = “Very Anxious”, 5 = “Worryingly Anxious”
- `Anxiety.Effects` Response to “What effect do you expect this course will have on your levels of anxiety towards statistics?” with response “Decrease”, “No effect”, “Increase” or “Don’t know”
- `Course` The course the sudent was ernolled on, i.e. “Bio”,“S1A” or“S1Y”
- `Time` When the questionnaire was completed: “Start” of course and “End” of course
- `SE.mean` Average response to the five “Self Efficacy” questions (see below)
- `SA.mean` Average response to the question ”Please indicate how much anxiety you would experience (from 1 = no anxiety, to 5 = strong anxiety) in each of the 23 situations (see below)
- `SA.Subscale1.mean` Average response to the first sub-scale of anxiety questions, namely “Test and class anxiety” (Q1, Q4, Q8, Q10, Q13, Q15, Q21, Q22)
- `SA.Subscale2.mean` Average response to the second sub-scale of anxiety questions, namely “Interpretation anxiety” (Q2, Q5, Q6, Q7, Q9, Q11, Q12, Q14, Q17, Q18, Q20)
- `SA.Subscale3.mean` Average response to the third sub-scale of anxiety questions, namely “Fear of asking for help” (Q3, Q16, Q19, Q23)
- `Exam` Final Exam resuls (as %)
- `GPA` Overall Grade Point Average including all continuous assessment (on 22 point scale)
- `Grade` Overall Grade including all continuous assessment, e.g. “A1”, “B2”, “C3”.

Five ”Self-Efficacy” questions (with responses 1 = “Not at all true”, 2= “Hardly true”, 3= “Moderately true” and 4=“Exactly true”)

1. I can always manage to solve difficult problems if I try hard enough.
2. It is easy for me to stick to my aims and accomplish my goals.
3. I can solve most problems if I invest the necessary effort.
4. I can remain calm when facing difficulties because I can rely on my coping abilities.
5. I can usually handle whatever comes my way.

Twenty-three ”Statistical Anxiety” scenarios (See Hanna et al, 2008)

1. Studying for an examination in a statistics course
2. Interpreting the meaning of a table
3. Going to ask my statistics teacher for individual help with material I am having difficulty understanding
4. Doing the coursework for a statistics course
5. Making an objective decision based on empirical data
6. Reading an article that includes some statistical analyses
7. Trying to decide which analysis is appropriate for a research project
8. Doing an examination in a statistics course
9. Reading an advertisement for a car which includes figures on miles per gallon, depreciation, etc.

10. Walking into the room to take a statistics test
11. Interpreting the meaning of a probability value
12. Arranging to have a dataset put into the computer
13. Finding that another student in class got a different answer than I did to a statistical problem
14. Determining whether to reject or retain the null hypothesis
15. Waking up in the morning on the day of a statistics test
16. Asking one of your lecturers for help in understanding output from an analysis
17. Trying to understand the odds in a lottery
18. Watching a student search through a load of computer output from his/her research
19. Asking someone in the computer lab for help in understanding output from an analysis
20. Trying to understand the statistical analyses described in the abstract of a journal article
21. Enrolling in a statistics course
22. Going over a final examination in statistics after it has been marked
23. Asking a fellow student for help in understanding output from an analysis

Question(s) of interest

Some questions of interest are:

1. What is the difference between statistics anxiety and self-efficacy between students of different courses, demographic and academic factors and achievement levels?

2. How do the students' levels of statistics anxiety change between the start and end of the course and is there a relationship between this change and performance?

3. What is the relationship between statistics anxiety and performance and course types, and how does self-efficacy and other variables affect this relationship?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Statistical Inference.
- Linear Models.
- Data Analysis Skills.
- Generalised linear models.

55 Cluster Analysis of Scottish Prescription Data (1)

Project level: Moderate

55.1 Overall project description

It is of interest to see if the pattern of GPs prescribing generic versus non-generic data follows the same overall trend across 12 months (February 2019 to January 2020) or if there are distinct groups which follow different patterns. Also of interest is how these groups and patterns relate to other information like the Health Board or area that the GP belongs to or the size of the practice (number of patients registered with them).

The methods suggested are clustering methods but functional data analysis and flexible regression techniques could also be used.

55.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data source: https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community#

The data (`PG.data.csv`) is made up of number of generic prescriptions (G) and non-generic prescriptions (P) of 1096 GP practices across the whole of Scotland. The data is monthly starting in February 2019 and continuing to January 2020.

An additional dataset (`meta.data.csv`) is availabe with some extra information about the practice size (`PracticeListSize`), the area of the GP (`GPCluster`), the datazone where the GP is located (`DataZone`) and the Healthboard the GP belongs to (`HB`).

Question(s) of interest

The main questions of interest are:

- What does the overall pattern of generic to non-generic prescription look like?
- What are the clusters that make up this distribution?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.

56 Modelling the Growth of Young Children (2)

Project level: Easy/Moderate

56.1 Overall project description

The study of human growth has long been of interest to scientists and health professionals. Understanding the factors and conditions that impact on the physical growth and development of children can help us make interventions which may improve the health of children in future generations. It is therefore critical that we have reliable methods that allow us to characterise a variety of different growth patterns in children.

For example, it would be extremely useful to be able identify and distinguish between children who are growing successfully and those whose growth is faltering. In the cases where children do falter, we wish to quantify the timing and the extent of their recovery. Once we have identified methodology for characterising growth patterns, it makes it much easier to begin to explore the factors which predict faltering and recovery.

In this project, you will have the opportunity to explore a real life example of infant growth data, and to compare different modelling approaches which can be used to develop growth curves for these children. Once you have selected a suitable model, you can use this to identify the differences between "healthy" and "unhealthy" growth and to consider some of the causes for these differences in growth.

56.2 Individual project details

How many individual projects are available in this area: 2.

56.3 Project 1

Data available

The *brokenstick* package in R includes a dataset containing longitudinal height and weight measurements for 206 Dutch children born in 1988 and 1989. This is a subset of a larger longitudinal study known as the SMOCC study, which contained 1933 children from across the Netherlands.

The subset of data for this project can be obtained using the *smocc.hgtwgt* function, and contains the following variables:

- `subjid` - The unique ID for each child in the study.
- `rec` - This outlines the record number for this child (for example, 3 means this is the third observation on that child).
- `nrec` - The total number of observations for the given child.
- `agedays` - The age of the child in days.

- sex - The sex of the child.
- ga - The gestational age of the child in days - ie the number of days that the child was in the womb before birth.
- bw - Birth weight in grams.
- htcm - Height measurement in cm.
- wtkg - Weight measurement in kg.

This project will focus on monitoring the **weight** of these children over time.

Question(s) of interest
The main questions of interest are:

- Does birthweight have an impact on subsequent growth in infants?
- Which types of models are most suitable for characterising growth patterns in children's weight?
- Are Z-scores better than raw data in terms of accurately modelling the growth of children (in terms of weight)?
- Can we develop a method for identifying the differences between "healthy" and "unhealthy" children (in terms of weight)?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Statistical Inference.

Knowledge of the following courses would be useful, but can be picked up during the project:

- Flexible Regression.
- Linear Mixed Models.
- Functional Data Analysis.

56.4 Project 2

Data available
The *brokenstick* package in R includes a dataset containing longitudinal height and weight measurements for 206 Dutch children born in 1988 and 1989. This is a subset of a larger longitudinal study known as the SMOCC study, which contained 1933 children from across the Netherlands.

The subset of data for this project can be obtained using the *smocc.hgtwgt* function, and contains the following variables:

- subjid - The unique ID for each child in the study.
- rec - This outlines the record number for this child (for example, 3 means this is the third observation on that child).
- nrec - The total number of observations for the given child.
- agedays - The age of the child in days.
- sex - The sex of the child.

- `ga` - The gestational age of the child in days - ie the number of days that the child was in the womb before birth.
- `bw` - Birth weight in grams.
- `htcm` - Height measurement in cm.
- `wtkg` - Weight measurement in kg.

This project will focus on monitoring the **height** of these children over time.

Question(s) of interest

The main questions of interest are:

- Does gestational age at birth have an impact on subsequent growth in infants?
- Which types of models are most suitable for characterising growth patterns in children's height?
- Are Z-scores better than raw data in terms of accurately modelling the growth of children (in terms of height)?
- Can we develop a method for identifying the differences between "healthy" and "unhealthy" children (in terms of height)?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Statistical Inference.

Knowledge of the following courses would be useful, but can be picked up during the project:

- Flexible Regression.
- Linear Mixed Models.
- Functional Data Analysis.

57 Quantifying pneumonia mortality risk in England (1)

Project level: Moderate

57.1 Overall project description

England is a country of around 55 million people, and pneumonia (International Classification of disease ICD-10 codes J12-J18) is a major cause of death across the country. This project aims to explore the spatio-temporal pattern in the population-level pnuemonia mortality risk in England, using data relating to the $K = 322$ local authorities that make up England at yearly intervals from 2002 to 2017 ($N = 16$ time periods). The aims of the analysis are to describe the key spatio-temporal dynamics in pnuemonia mortality risk, including: (i) identifying potential risk factors that influence pnuemonia mortality risk; (ii) estimating the temporal trend in pnuemonia mortality risk; and (iii) quantifying the magnitude of the spatial variation in pnuemonia mortality risk between local authorities which measures the level of inequality in this disease.

57.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available about pneumonia mortality incidence, population demography, and covariates, including air pollution concentrations and poverty, for each of the $K = 322$ local authorities for each year between 2002 to 2017. The data are stored in `EnglandLUAdata.csv` and contain the following columns.

- `Code` - A unique code for the local authority area.
- `Name` - The name of the local authority area.
- `Year` - The year the data relate to.
- `Y` - The number of pnuemonia mortalities in each local authority in the year.
- `E` - The expected number of pnuemonia mortalities in each local authority in the year based on the population size and demographic structure of the people who live in the local authority. This is computed by indirect standarisation, for details see chapter 1 of the Biostatistics course.
- `PM25` - Measure of fine particulate matter air pollution.
- `IMD` - The English Index of Multiple Deprivation (IMD), where higher values indicate increasing levels of poverty.

There are also shapefiles available giving the local authority boundaries, which are a collection of files with the name `LocalAuthorities` with different file extensions.

Question(s) of interest

The main questions of interest are:

- What effects do the air pollution and socio-economic covariates have on disease risk?
- What is the temporal trend in pneumonia mortality risk and which local authorities exhibit the highest risks?
- How big are the inequalities (spatial variations) in pneumonia mortality risk across England, and are these inequalities increasing or decreasing over time?

Relevant courses

We recommend that you have taken the following courses to undertake this project:

- Generalised linear models (strongly recommended).
- Spatial statistics (desirable).

58 Spatio-temporal analysis of sea surface temperatures (1)

Project level: Moderate/Difficult

58.1 Overall project description

West-blowing trade winds in the Indian Ocean push warm surface waters against the eastern coast of Africa. These waters move south along the coastline, eventually spilling out along the boundary of the Indian and Atlantic Oceans. This jet of warm water, known as the Agulhas Current, collides with the cold, west to east owing Antarctic Circumpolar Current, producing a dynamic series of meanders and eddies as the two waters mix. The result makes for an interesting target for spatio-temporal analysis. In this project, you will use sea surface temperature (SST) data collected by satellite for the Agulhas and surrounding areas off the coast of South Africa to fit spatio-temporal models with a view towards prediction of missing values.

58.2 Individual project details

How many individual projects are available in this area: 1.

Data available
The data `SSTagulhas.mat` contains sea surface temperature data collected by satellite for the Agulhas and surrounding areas off the coast of South Africa from January 1 to November 26, 2004, a period of 331 days. The main variable, `SST.zone.period`, is a three dimensional 72 x 240 x 331 matrix (latitude, longitude, day) of sea surface temperatures given in degrees Celsius. Spatial resolution for the data set is roughly 25 kilometres, though exact values depend on latitude. Longitude and latitude values are respectively stored in `lon.zone` and `lat.zone`. Temporal resolution is one day.

Question(s) of interest
The main questions of interest are:

- Do SST vary over space?
- Do SST vary over time?
- Do universal kriging models can help us in successfully predict missing values in a given day?
- Can a spatio-temporal kriging approach produce better predictions?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Spatial statistics and Time series.

59 Measles Susceptibility in Scotland (1)

Project level: Easy/Moderate

59.1 Overall project description

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 since 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.

59.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available for children eligible to attend pre-school from the SCIRS database at intermediate zone (IZ) level in Scotland, which are small geographical units containing, on average, 4000 residents between 1998 and 2014. The data are stored in `MMR.csv` and contain the following columns.

- `Year` - The year the data relate to.

- `IZ` - A unique intermediate zone code for each area in Scotland.
- `Total_child` - The total number of pre-school children in the specified IZ and year.
- `Num_Susc` - The number of pre-school children susceptible to measles in the specified IZ and year.

Two further files, `Scotland study area.shp` and `Scotland study area.dbf`, are also available and can be used to plot maps of the study region. To work with these shapefiles you will need the `shapefiles`, `spdep` and `CARBayes` libraries in `R`.

Question(s) of interest

The main questions of interest are:

- What impact did the controversy surrounding the Wakefield paper have on the overall temporal trend in vaccination rates in Scotland?
- Did the magnitude of the spatial inequality in measles susceptibility in Scotland change due to the MMR scare?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

60 Modelling biological systems with nonparametric models (1)

Project level: Moderate/Difficult

60.1 Overall project description

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 smoothing splines model and compare it with some of the more well-known nonparametric model such as the Reproducing Kernel Hilbert Space (RKHS) regression model.

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

60.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data can be generated from the R package KGode by running the code below. The ODE (ordinary differential equation) model used here is the Lotka-Volterra model (prey-predator). From the simulated dataset below, we have y_no, the noised observation of the two states variables in the ODE model. And t_no, the time points for each observation.

```r
require(mvtnorm)
library(KGode)

noise = 0.1  ## set the variance of noise
SEED = 19537
set.seed(SEED)

## Define ode function, we use lotka-volterra model in this example.
## we have two ode states x[1], x[2] and four ode parameters alpha, beta,
## gamma and delta.
LV_fun = function(t,x,par_ode){
  alpha=par_ode[1]
  beta=par_ode[2]
  gamma=par_ode[3]
  delta=par_ode[4]
  as.matrix( c( alpha*x[1]-beta*x[2]*x[1] , -gamma*x[2]+delta*x[1]*x[2] ) )
}

## create a ode class object
kkk0 = ode$new(2,fun=LV_fun)
```

```
## ode is sample 2.
```

```r
## set the initial values for each state at time zero.
xinit = as.matrix(c(0.5,1))
## set the time interval for the ode numerical solver.
tinterv = c(0,6)
## solve the ode numerically using alpha=1, beta=1, gamma=4, delta=1.
kkk0$solve_ode(c(1,1,4,1),xinit,tinterv)
## Add noise to the numerical solution of the ode model and
## treat it as the noisy observation.
y_true= t(kkk0$y_ode)
n_o = max( dim( y_true) )
t_no = kkk0$t
y_no = y_true + rmvnorm(n_o,c(0,0),noise*diag(2))
```

Students are also encrouaged to try other ODE models such as the SIR model which is widely used in epidemiology.

Question(s) of interest

The main questions of interest are:

- How does the smooth spline model perform in comparison with the more commonly used parametric models such as linear regression models?
- How does the smooth spline model perform in comparison with the other nonparametric models such as RKHS?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Linear regression models.
- Flexible regression.

61 Model selection: review and simulation (1)

Project level: Easy/Moderate

61.1 Overall project description

One common task in data analysis is to identify one model of a set of data out of some discrete set of models as the best of that class. As an example, think of a multiple regression problem with p potential covariates. Here, just considering models with no interactions, there are 2^p possible models (why?). Which one is best? (In this context, model selection is often called variable selection: i.e., which variables matter?)

There is a huge number of approaches to this problem. In this project, first you should review some of those that you have already seen and find out about at least one other approach that is new to you. You should describe how the approaches work, what principles they are based upon and compare approaches as far as possible. You might find it easiest to think of the problem in the specific regression context above, where you have n observations of the response and their corresponding sets of potentially p covariate values, and where the biggest model is

$$Y_i = \alpha + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + e_i.$$

Here Y_i is the response in the ith sample, $(x_{i,1}, \dots, x_{i,p})$ are the values of the covariates in that sample, the error term of which is $e_i \sim N(0, \sigma^2)$, independently across samples, so that σ^2 is the (unknown) variance of the errors. The $(\beta_1, \dots, \beta_p)$ are the regression coefficients, many of which might be equal to zero. Let there be q of the β_is which are non-zero. The model selection task here is to find the q covariates with those non-zero β_is.

There are many methods to do this and many ways to review them. You certainly do not have time to cover them all. Below (in Questions of Interest) are some ideas of things you might want to address.

The second part of your project should use a simulation approach to explore some comparison(s) of interest. The idea is to choose a multiple regression model including only certain covariates, repeatedly simulate values of the full set of potential covariates and the corresponding error terms and, from them, simulate the values of the response. Then plug this simulated data into certain model selection methods and compare how they perform, for example, at finding the correct model.

61.2 Individual project details

How many individual projects are available in this area: 1.

Data available

This is a literature and simulation-based project.

Questions of interest
Possible questions of interest include:

- What different approaches to model selection can be taken?
- Can all models be compared? If not, what techniques are available?
- How do approaches differ?
  - What are distinct criteria used to find good models (e.g., adjusted R^2, Mallows' C_p, AIC, BIC, DIC, …)?
  - You might contrast approaches that are based on statistical significance (e.g., F tests) to those based on achieving the best out-of-sample prediction performance (mean-square error).
  - You might contrast a Bayesian approach (e.g., based on Bayes' factors) to a frequentist approach based on statistical significance.
  - You probably want to distinguish carefully the criterion that is used to determine the best model (e.g., nested F tests) from the algorithm used to try to find it (e.g., stepwise selection).
  - You might explain how penalized regression can perform model selection (e.g., the Lasso or ridge regression).
- Of two (or more) approaches that you choose, which of them does better on simulated data at finding the model from which the data was simulated?
- How does that depend on the noise level σ^2?
- Of those approaches, which does better at predicting out-of-sample data, e.g., in terms of mean-square error?

Note: these are suggestions, but you can address other questions related to the topic, so long as you review the background methodology and/or explore them by simulation.

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Statistical Inference.
- Regression Modelling.
- Bayesian Statistics.
- Machine Learning.

62 Temporal clustering of ecological indices (1)

Project level: Moderate

62.1 Overall project description

Ecological time series are commonly available data counting the number of species available at a location for specific time points. It is often interesting to look at similarities in these indices across space and/or species to try and determine possible drivers of the observed trends. This project will investigate methods for clustering time series of robins and starlings observed over mutliple locations.

62.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on numbers of annual averaged counts of robins and starlings, across 693 monitoring sites, spanning 45 winters. There is also a selection of environmental and spatial variables for the same location and year. The data are stored in `Garden_bird.csv` and contain the following columns.

- `site` - A unique code for each garden.
- `year` - The year of observation (year 1 being the winter of 1970/71)
- `subrur` - A two-level variable stating whether the site is (sub)urban (+1) or rural (-1).
- `easting` - A standardised longitude measure.
- `northing` - A standardised lattitude measure.
- `temp,frost` - Measures of average temperature/number of days of ground frost respectively over the winter months.
- `sparr` - Averaged number of sparrowhawks (a predator) observed, considered by some to be a potential important driver of population changes.
- `robin,starl` - The response variables (averaged counts of robins and starlings respectively).

Questions of interest

The main questions of interest are:

- Can we cluster the different garden time-series into similar groups of gardens using time series clustering methods?
- Are there differences in the underlying properties of the gardens in each cluster?
- Can we predict which diffferences, if any?

Relevant courses

We recommend that you have taken the following courses to undertake this project:

- Multivariate Methods.
- Time Series.
- Linear models/Regression.

63 Species richness, stress and connectivity in UK lakes (2)

Project level: Moderate

63.1 Overall project description

Hydroscape (https://hydroscapeblog.wordpress.com/) is a NERC-funded project aiming to determine how stressors and connectivity interact to influence biodiversity and ecosystem function in freshwaters across the UK. Connectivity between freshwaters is a major factor behind dispersal of both stressors and biodiversity, but how the effects of stressors and connectivity interact to influence species richness is poorly understood. The aim of these projects are to model the effects of stress, connectivity, and their interactions, on species richness of beetles and molluscs.

The datasets have been made available as part of the Hydroscape project. The response data were compiled by Dr. Alan Law of the University of Stirling and the explanatory variables were compiled by Dr. Philip Taylor of the UK Centre for Ecology and Hydrology.

63.2 Individual project details

How many individual projects are available in this area: 2.

63.3 Project 1 - Are beetles more stressed in connected freshwaters?

Data available
A dataset named `beetles_project_data.csv` contains the following columns:

- `GridRef1km`: OSGB 1km grid reference of lake (This is only of use if you wish to calculate the longitude and latitude of each lake, e.g. for plotting purposes, via the function `osg_parse` in the `rnrfa` R package.)
- `nSpecies`: Beetle species richness (the response variable)
- `WBID`: UK Lakes portal water body ID (This is only of use if you wish to look up a particular lake, e.g. in there are unusual observations like a very large lake area.)
- `lake_data_Lake.area..ha.`: Lake area (ha) (t)
- `lake_data_Lake.altitude..m.`: Lake altitude (m above mean sea level) (t)
- `lake_data_Mean.Depth..m.`: Lake mean depth (m) (t)
- `lake_data_Volume..m3.`: Lake volume (m^3) (t)
- `lake_data_Perimeter..km.`: Lake perimeter (m) (t)
- `lake_data_WFD_River_Basin_District`: Water Framework Directive River Basin District (t)

- `landscape_2km_Mean.slope..degrees.`: Mean slope of 2km buffer around lake (c)
- `landscape_2km_Lake.area..`: % of 2km buffer around lake covered by lakes (c)
- `landscape_2km_Pond.area..`: % of 2km buffer around lake covered by ponds (c)
- `landscape_2km_Rivers...length..m._per_ha`: Total length of rivers in 2km buffer around lake (c)
- `landscape_2km_Canals...length..m._per_ha`: Total length of canals in 2km buffer around lake (c)
- `landscape_2km_Obstacles...Count_per_ha`: Total number of obstacles in 2km buffer around lake (c)
- `landscape_2km_Lakes...Perimeter..m._per_ha`: Total perimeter of all other lakes in 2km buffer around lake (c)
- `landscape_2km_Ponds...Perimeter..m._per_ha`: Total perimeter of all ponds in 2km buffer around lake (c)
- `landscape_2km_Lakes...Count_per_ha`: Total number of other lakes in 2km buffer around lake (c)
- `landscape_2km_Ponds...Count_per_ha`: Total number of ponds in 2km buffer around lake (c)
- `landscape_2km_LCM2007...Agricultural..`: % agricultural land cover in 2km buffer around lake (s)
- `landscape_2km_LCM2007...Urban..`: % urban land cover in 2km buffer around lake (s)
- `landscape_2km_Mean.Temperature..2000.2016..C.`: mean temperature of 2km buffer around lake (c)
- `landscape_2km_Mean.Annual.Rainfall..2000.2016..mm.`: mean annual rainfall of 2km buffer around lake (c)
- `landscape_2km_Visits...Fishing`: Normalised measure of number of visits for fishing purposes in 2km buffer around lake (higher = more visits) (c)
- `landscape_2km_Visits...Watersports`: Normalised measure of number of visits for watersport purposes in 2km buffer around lake (higher = more visits) (c)
- `landscape_2km_Visits...All`: Normalised measure of number of visits for all purposes in 2km buffer around lake (higher = more visits) (c)

The variables marked (c) are the connectivity variables, those marked (s) are the stressors and those marked (t) are lake typology variables that may be useful to control for in any model.

Questions of interest
The main questions of interest are:

- What are the main relationships between species richness and stress due to land cover surrounding freshwaters?
- What are the main relationships between species richness in freshwaters and connectivity to other nearby freshwaters?
- Does the relationship between species distribution and stressors change depending on connectivity to other nearby freshwaters?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Flexible regression.

63.4 Project 2 - Are molluscs more stressed in connected freshwaters?

Data available
A dataset named `molluscs_project_data.csv` contains the following columns:

- `GridReference1km`: OSGB 1km grid reference of lake (This is only of use if you wish to calculate the longitude and latitude of each lake, e.g. for plotting purposes, via the function `osg_parse` in the `rnrfa` R package.)
- `nSpecies`: Beetle species richness (the response variable)
- `WBID`: UK Lakes portal water body ID (This is only of use if you wish to look up a particular lake, e.g. in there are unusual observations like a very large lake area.)
- `lake_data_Lake.area..ha.`: Lake area (ha) (t)
- `lake_data_Lake.altitude..m.`: Lake altitude (m above mean sea level) (t)
- `lake_data_Mean.Depth..m.`: Lake mean depth (m) (t)
- `lake_data_Volume..m3.`: Lake volume (m^3) (t)
- `lake_data_Perimeter..km.`: Lake perimeter (m) (t)
- `lake_data_WFD_River_Basin_District`: Water Framework Directive River Basin District (t)
- `landscape_2km_Mean.slope..degrees.`: Mean slope of 2km buffer around lake (c)
- `landscape_2km_Lake.area..`: % of 2km buffer around lake covered by lakes (c)
- `landscape_2km_Pond.area..`: % of 2km buffer around lake covered by ponds (c)
- `landscape_2km_Rivers...length..m._per_ha`: Total length of rivers in 2km buffer around lake (c)
- `landscape_2km_Canals...length..m._per_ha`: Total length of canals in 2km buffer around lake (c)
- `landscape_2km_Obstacles...Count_per_ha`: Total number of obstacles in 2km buffer around lake (c)
- `landscape_2km_Lakes...Perimeter..m._per_ha`: Total perimeter of all other lakes in 2km buffer around lake (c)
- `landscape_2km_Ponds...Perimeter..m._per_ha`: Total perimeter of all ponds in 2km buffer around lake (c)
- `landscape_2km_Lakes...Count_per_ha`: Total number of other lakes in 2km buffer around lake (c)
- `landscape_2km_Ponds...Count_per_ha`: Total number of ponds in 2km buffer around lake (c)
- `landscape_2km_LCM2007...Agricultural..`: % agricultural land cover in 2km buffer around lake (s)
- `landscape_2km_LCM2007...Urban..`: % urban land cover in 2km buffer around lake (s)

- `landscape_2km_Mean.Temperature..2000.2016..C.`: mean temperature of 2km buffer around lake (c)
- `landscape_2km_Mean.Annual.Rainfall..2000.2016..mm.`: mean annual rainfall of 2km buffer around lake (c)
- `landscape_2km_Visits...Fishing`: Normalised measure of number of visits for fishing purposes in 2km buffer around lake (higher = more visits) (c)
- `landscape_2km_Visits...Watersports`: Normalised measure of number of visits for watersport purposes in 2km buffer around lake (higher = more visits) (c)
- `landscape_2km_Visits...All`: Normalised measure of number of visits for all purposes in 2km buffer around lake (higher = more visits) (c)

The variables marked (c) are the connectivity variables, those marked (s) are the stressors and those marked (t) are lake typology variables that may be useful to control for in any model.

Questions of interest

The main questions of interest are:

- What are the main relationships between species richness and stress due to land cover surrounding freshwaters?
- What are the main relationships between species richness in freshwaters and connectivity to other nearby freshwaters?
- Does the relationship between species distribution and stressors change depending on connectivity to other nearby freshwaters?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Flexible regression.

64 COVID-19 in Bangladesh (1)

Project level: Moderate

64.1 Overall project description

The COVID-19 pandemic reached Bangladesh in early 2020, with the first case in March 2020, before spreading throughout the country. Daily tests and confirmed case numbers are recorded by Bangladesh's Directorate General of Health Services, for each of the 64 districts of Bangladesh. The main aim of this project is to investigate the patterns of cases over space between October 2020 and April 2021, and whether they relate to the available covariates.

64.2 Individual project details

How many individual projects are available in this area: 1.

Data available

A shapefile of the 64 districts of Bangladesh is stored in the shapefile named `BGD_districts.shp`. The `data` slot of the shapefile contains the following columns:

- `DISTNAME`: Name of each district
- `DISTCODE`: A numeric code for each district

The weekly case and test numbers for each district are stored in the CSV file named `BGD_cases.csv`, along with covariate information. The file contains the following columns:

- `DISTNAME`: Name of each district (s)
- `week`: End date for each week (t)
- `cases`: Weekly confirmed case numbers for each district (st)
- `tests`: Weekly tests for each district (st)
- `popcount`: Population count per district (s)
- `popdens`: Population density per district (s)
- `health_travel_time_walk`: Mean walking travel time to healthcare per district (s)
- `health_travel_time_motor`: Mean motorised travel time to healthcare per district (s)
- `dens_road`: Road density per district (s)
- `dens_rail`: Rail density per district (s)
- `wealth`: Wealth score per district (s)
- `income`: Mean income per district (s)
- `urba_percent`: Urban landcover percentage per district (s)
- `wetland_percent`: Wetland landcover percentage per district (s)
- `rural_percent`: Rural landcover percentage per district (s)
- `temp`: Mean temperature per district and per week (st)
- `precip`: Mean rainfall per district and per week (st)

- `mob.residential`: Google mobility score for residential per week (t)
- `mob.pubtport`: Google mobility score for public transport per week (t)
- `mob.retailrec`: Google mobility score for retail and recreation per week (t)
- `mob.workplace`: Google mobility score for workplace per week (t)
- `mob.parks`: Google mobility score for parks per week (t)
- `mob.grocpharm`: Google mobility score for groceries and pharmacies per week (t)

The variables marked (s) are spatial only, those marked (t) are temporal only and those marked (st) are spatiotemporal (i.e. take values for each district and for each week).

The temporal and spatiotemporal covariates (in columns 16 to 23) are each lagged by one week.

Questions of interest
The main questions of interest are:

- What are the main spatial and temporal patterns in the data?
- At the peak of the wave (i.e. for the week with the highest case numbers ending 2021-04-08), how do the relevant covariates relate to the cases across space?
- How do the covariates relate to the case numbers over space and time?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Time series.
- Spatial statistics.

65 A spatio-temporal approach to identify occurrence areas and abundance hotspots of the European hake (1)

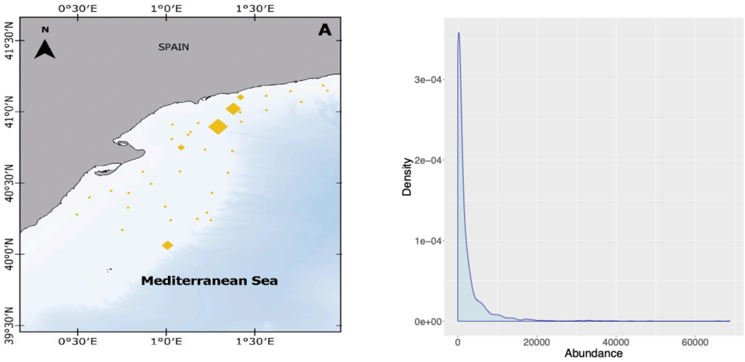
Project level: Moderate

65.1 Overall project description

We are interested in describing the occurrence and abundance of the European hake in wintertime. The study area is the north-western Mediterranean Sea, between the 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 km2 and a depth range from 30 to 350 m, including the continental shelf and upper slope. Data on hake location and abundance were collected as part of a trawl survey project between 1994 and 2017. Sampling stations were placed randomly within each bathymetric stratum at the beginning of the project. Sampling was performed in similar geographical locations in all subsequent years.

Our response variable is hake abundance, 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 some extreme values.

We want to develop two models that adequately capture 1) the presence/absence and 2) the abundance conditional to presence. This will allow us to identify high occurrence areas and highlight abundance hotspots on the same spatial scale. Each model should incorporate different spatial, temporal or spatio-temporal effects and information about environmental and geographical factors. To this end, we will use a generalised additive model (GAM) framework.



(a)

(b)

Figure 1: Study area (a) and kernel density estimate (b) for hake abundance.

65.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data `hake.txt` contain 1971 observations and 6 variables. The variables are

- `year`: year of the trawling.
- `haul`: points in the ocean where trawling is taking place. In oceanographic campaigns, it is done for a standard time of 30 min so that they are all comparable. For our data, these locations correspond to the yellow diamond in the map shown above.
- `lon`, `lat`: longitude and latitude.
- `abundance`: hake abundance (our response variable).
- `biomass`: is the weight of the fish caught and is measured as kg every 30 min of trawling.

Additional covariates such as the **distance to the coast** can be constructed using, e.g., the function `getNOA.bathy` from the `marmap R` package.

Modelling framework

1. **Model 1: presence/absence.** Let $Y(s,t)$ be the occurrence (1 being yes; 0 being no) of hake for each haul at location $s \in \mathcal{S}$ in year $t \in \mathcal{T}$, where $\mathcal{S}$ is the study area and $\mathcal{T} = \{1994, \dots, 2017\}$ is the set of all available years. Then, $Y(s,t)$ can be modeled as

$$\begin{aligned} Y(s,t) &\sim \text{Bernoulli}(\pi(s,t)), \quad s \in \mathcal{S}, t \in \mathcal{T}, \\ \text{logit}(\pi(s,t)) &= \beta_{0,\pi} + \beta_{1,\pi}\texttt{biomass}_{s,t} + f_1(\texttt{lat}_s, \texttt{lon}_s) + f_2(t), \end{aligned} \tag{1}$$

   where $\texttt{lat}_s$ and $\texttt{lon}_s$ are the latitude and longitude of the site s. Within the GAM framework, the unknown functions f_1 and f_2 can be represented using reduce rank smoothing splines (more specifically, the reduced rank isotropic thin plate splines; see Wood, 2017, Chapter 4). The smooth interaction term f_1 is constructed using the tensor product construction where the main effects of `lat` and `lon` are excluded (Wood, 2016). The significance of each term in (1) should be assessed.

2. **Model 2: hake abundance.** Let $Z(s,t)$ be the positive abundance for each haul at location $s \in \mathcal{S}$ in year $t \in \mathcal{T}$. Then, $Z(s,t)$ can be modeled using a location-scale Gamma model as follows:

$$\begin{aligned} Z(s,t) &\sim \text{Gamma}(\mu(s,t), \sigma(s,t)), \quad s \in \mathcal{S}, t \in \mathcal{T} \\ \mu(s,t) &= \beta_{0,\mu} + \beta_{1,\mu}\texttt{biomass}_{s,t} + f_3(\texttt{lat}_s, \texttt{lon}_s) + f_4(t), \\ \log\{\sigma(s,t)\} &= \beta_{0,\sigma} + \beta_{1,\sigma}\texttt{biomass}_{s,t} + f_5(\texttt{lat}_s, \texttt{lon}_s) + f_6(t), \end{aligned} \tag{2}$$

   where the functions $f_3, \dots, f_6$ have the same interpretation as in (1). The Gamma likelihood is parametrised in terms of its mean $\mu(s,t)$ and its standard deviation $\sigma(s,t)$. We can retrieve the usual rate (β) and shape (α) Gamma parametrisation using the fact that $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$.

3. **Model 3: improving the tail of hake abundance.** The Gamma distribution has light tails and might not be suitable to capture extreme abundance. We want to investigate if a generalised Pareto distribution fitted to large abundances (*abundance exceedances*) does a better job. To this end, we assume that for some large threshold $u(s,t)$, the distribution of hake abundance larger than u follows a generalised Pareto (GP) distribution, i.e.,

   $$\mathrm{P}(Z(s,t) - u(s,t) > z \mid Z(s,t) > u(s,t)) = \left(1 + \frac{z \cdot \xi(s,t)}{\tilde{\sigma}(s,t)}\right)^{-1/\xi(s,t)},$$

   where ξ is the tail (or shape) parameter, and $\tilde{\sigma}(s,t)$ is the scale parameter. Functional forms for the GP parameters can be constructed following formulae in (2). Then, the model can be expressed as

   $$\begin{aligned} Z(s,t) - u(s,t) > z \mid Z(s,t) > u(s,t) \sim \mathrm{GP}(\xi(s,t), \hat{\sigma}(s,t)), \quad s \in \mathcal{S}, t \in \mathcal{T}, \\ \xi(s,t) = \beta_{0,\xi} + \beta_{1,\xi}\texttt{biomass}_{s,t} + f_7(\texttt{lat}_s, \texttt{lon}_s) + f_8(t), \\ \log\{\sigma(s,t)\} = \beta_{0,\tilde{\sigma}} + \beta_{1,\tilde{\sigma}}\texttt{biomass}_{s,t} + f_9(\texttt{lat}_s, \texttt{lon}_s) + f_{10}(t), \end{aligned}$$

   Question(s) of interest
   The main questions of interest are:

- Understanding the distributional patterns associated with hake presence and abundance.
- Is there any significant spatio-temporal trends in hake abundance?
- Is there any significant spatio-temporal trends in the presence/absence model?
- Is there any significant improvement by using the GP distribution for abundance exceedances?
- Are the bulk and the tail of the hake abundance distribution driven by the same effects, i.e., are they all affected by time, space, or any other covariate?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Advanced predictive models.
- Spatial statistics.

66 What are the drivers of Covid-19 mortality in England? (2)

Project level: Moderate

66.1 Overall project description

The Covid-19 pandemic has caused worldwide devastation throughout 2020 and into 2021, and by the end of 2020 was in its second wave of infection in England. The first wave happened between March and August (inclusive) 2020, while the second wave started in September 2020 and for the purposes of this study ends at the end of January 2021 (the last month for which we have data). The aim of this project is to undertake a study quantifying the drivers and hotspots of elevated Covid-19 mortality rates in England at the small-area scale, as well as examining how the spatial distributions of the mortality rates differ between the two waves of the pandemic. The study uses routinely collected small-area Covid-19 mortality data and associated risk factors for the K = 6,772 Middle Super Output Areas (MSOA) in mainland England, and covers both the first and second waves of the pandemic between March 2020 and Januray 2021.

66.2 Individual project details

How many individual projects are available in this area: 2.

66.3 Project 1 - The drivers and spatial patterns in raw Covid-19 mortality rates

Data available

Data are available on Covid-19 mortality rates and its possible drivers for each MSOA in mainland England, and are stored in `Covid19MSOAdata project1.csv` and contain the following columns.

- `MSOA` - Unique code for each MSOA.
- `name` - Name of each MSOA
- `rawrate_all, rawrate_wave1, rawrate_wave2` - Raw mortality rates due to Covid 19 (that is the number of deaths per 1,000 people) for both waves of the pandemic and each of the two waves separately.
- `popdens` - The population density, i.e. the number of people per hectare.
- `no2, pm10, pm25` - Air pollution concentrations, specifically nitrogen dioxide (no2), particulate matter less that 10 microns in diameter (pm10) and particulate matter less that 2.5 microns in diameter (pm25).
- `Total.income` - Average total income.
- `Net.income` - Average net income after tax.

- `imdoverall, imdincome, imdemployment, imdeducation, imdcrime, imdhousing, imdenvironment` - Measures of socio-economic deprivation called the Index of Multiple Deprivation (IMD). The overall variable is the final composite index, while the remaining variables relate to specific domains of poverty. High values denote poorer areas.
- `region` - The NHS region of England the MSOA is within.
- `temperature` - Yearly average temperature levels.
- `cases_rate_msoa` - The number of confirmed Covid-19 cases per 1,000 of the population.
- `perc.male` - The percentage of the population who are male.
- `perc.15.44, perc45.54, perc55.64, perc65.74, perc75.84, perc85.89, perc90plus` - The percentages of the population in each age group.
- `percwhite, percblack, percindpakban, percchina` - The percentage of the population who are of different ethnicities (white, black, Indian/Pakistani/Bangladeshi, Chinese).
- `carehome` - The number of care homes.

Questions of interest
The main questions of interest are:

- What factors are associated with elevated or reduced Covid-19 mortality rates?
- Do these factors vary by the wave of the pandemic?
- Are the highest Covid-19 mortality rates in the same MSOAs in each wave of the pandemic?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Regression models.
- Spatial statistics (possibly, depending on which way you take the project).

66.4 Project 2 - The drivers and spatial patterns in relative Covid-19 mortality rates

Data available
Data are available on Covid-19 mortality rates and possible its drivers for each MSOA in mainland England, and are stored in `Covid19MSOAdata project2.csv` and contain the following columns.

- `MSOA` - Unique code for each MSOA.
- `name` - Name of each MSOA
- `relativerate_all, relativerate_wave1, relativerate_wave2` - Relative mortality rates due to Covid 19 for both waves of the pandemic and each of the two waves separately. These relative rates are relative to the average mortality rates in England, and have been adjusted to allow for the size and the age/sex demographics of the populations in each MSOA. Here a relative rate of 1 corresponds to the England-wide national rate, while a rate of 1.2 corresponds to a 20% increased rate of mortality com-

pared to the English average. Similarly, a rate of 0.9 corresponds to a 10% decreased rate of Covid-19 mortality compared to the English national average.
- `popdens` - The population density, i.e. the number of people per hectare.
- `no2, pm10, pm25` - Air pollution concentrations, specifically nitrogen dioxide (no2), particulate matter less that 10 microns in diameter (pm10) and particulate matter less that 2.5 microns in diameter (pm25).
- `Total.income` - Average total income.
- `Net.income` - Average net income after tax.
- `imdoverall, imdincome, imdemployment, imdeducation, imdcrime, imdhousing, imdenvironment` - Measures of socio-economic deprivation called the Index of Multiple Deprivation (IMD). The overall variable is the final composite index, while the remaining variables relate to specific domains of poverty. High values denote poorer areas.
- `region` - The NHS region of England the MSOA is within.
- `temperature` - Yearly average temperature levels.
- `cases_rate_msoa` - The number of confirmed Covid-19 cases per 1,000 of the population.
- `percwhite, percblack, percindpakban, percchina` - The percentage of the population who are of different ethnicities (white, black, Indian/Pakistani/Bangladeshi, Chinese).
- `carehome` - The number of care homes.

Questions of interest

The main questions of interest are:

- What factors are associated with elevated or reduced Covid-19 mortality rates?
- Do these factors vary by the wave of the pandemic?
- Are the highest Covid-19 mortality rates in the same MSOAs in each wave of the pandemic?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Regression models.
- Spatial statistics (possibly, depending on which way you take the project).

67 Estimating changes in health inequalities in respiratory disease across Scotland (1)

Project level: Moderate

67.1 Overall project description

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.

67.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The dataset contains counts of the number of hospital admissions for respiratory disease in each intermediate zone (IZ) in Scotland from 2003 - 2012. The country is divided into 1235 IZs which are non-overlapping small administrative areas, which contain on average 4000 household residents. The dataset also contains a number of covariates relating to each region.

This data was obtained from the Scottish government via https://statistics.gov.scot/home. Students are welcome to download additional data from this website in order to improve their model, but this is entirely optional.

The data are stored in `RespiratoryDisease.csv` and contains the following columns.

- `IZ` - The intermediate zone code.
- `Ei` - The expected number of hospital admissions for each IZ.
- `Yi` - The observed number of hospital admissions for each IZ.

- `JSA, Urban, Percent.Asian, Percent.Black` - Selected covariates for each zone - these include the percentage of 16-64's claiming job seekers allowance for each IZ, whether an area is deemed to be urban (1) or rural (2) (search Scottish Government Urban Rural Classification for more details), the proportion of people of Asian ethnicity in each IZ and the proportion of people of Black ethnicity in each IZ.
- `Year` - The year which the hospital admissions relate to.

Question(s) of interest

The main questions of interest are:

- 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?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Time Series.
- Spatial statistics.

68 Hospital admissions due to respiratory disease in Greater Glasgow and Clyde (1)

Project level: Easy/Moderate

68.1 Overall project description

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). 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.

68.2 Individual project details

How many individual projects are available in this area: 1.

Data available

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 data are stored in `respiratory admissions.csv` and `respiratory admissions expected counts.csv` and contain the following columns.

- `IG` - A unique intermediate geography code for each area in Scotland.
- `Y2002`,…,`Y2011` - The number of hospital admissions due to respiratory disease in the given year.
- `E2002`,…,`E2011` - The number of expected hospital admissions due to respiratory disease in the given year.

Two further files, `Scotland study area.shp`, `Scotland study area.dbf`, and `GlasgowIG.csv` are also available and can be used to plot maps of the study region. To work with these shapefiles you will need the `shapefiles`, `spdep` and `CARBayes` libraries in `R`.

Question(s) of interest

The main questions of interest are:

- Which areas exhibit an increase, a decrease, or no change in risk over the 10-year period?
- How have these changes in risk impacted upon health inequalities?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models.
- Spatial statistics.

69 Investigating the G-Base heavy metal soil content in Glasgow (1)

Project level: Moderate

69.1 Overall project description

The geochemical content of soil has the potential to adversely affect human, animal, and plant health. Heavy metals such as Lead (Pb), a common component of contaminated soils, are particularly harmful when found in higher concentrations. Urban areas with a rich industrial heritage, such as the Glasgow conurbation in the Clyde River Basin, are more likely to contain pockets of high soil contamination either as a direct result of industrial and other waste disposal or indirectly as an effect of high population density. The main aim of the study is to investigate the processes that leads to the accumulation of Pb in the topsoil and identify areas of soil contamination. The study uses the G-BASE dataset, collected by the British Geological Survey (BGS), along with covariates known to have an effect on heavy metal dispersion and soil variation.

69.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on Pb concentration in parts per million (ppm) are available throught the G-BASE dataset. Other environmental variables are also provided. The data are stored in `GbaseProjectPb.csv` and contain the following columns

- `X` - Sample location - Eastings (British National Grid)
- `Y` - Sample location - Northings (British National Grid)
- `Pb` - Lead concentration in parts per million (ppm)
- `Elevation` - Sample elevation in meters (m)
- `Slope` - Sample slope
- `Aspect` - Sample aspect
- `Plan.Curvature` - Sample plan curvature
- `Profile.Curvature` - Sample profile curvature
- `TWI` - Topographic Wetness Index
- `MRVBF` - Multiresolution Index of Valley Bottom Flatness
- `MRRTF` - Multiresolution Ridge Top Flatness
- `Population` - Population density per km^2
- `Landuse` - 21 distinct landuse categories

Questions of interest

The main questions of interest are:

- What is the spatial pattern of Pb contamination in the Glasgow area?
- Are there clusters or hotspots of contamination?
- What is the effect of the covariates on the Pb spatial pattern?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Linear Models.
- Spatial Statistics.
- Environmental Statistics.

70 Abundances of seabirds and identification of common trends (1)

Project level: Moderate

70.1 Overall project description

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.

These data have been provided by JNCC and should be fully acknowledged.

70.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data are stored in `SMP seabird abundance data 1986-2018.xlsx` and contain the following columns.

- `Country` - Country (Scotland, etc.)
- `County` - County (Norfolk, etc.)
- `MasterSite` - a larger area that contains at least one colony/site
- `Site` - colony/site within a MasterSite
- `SiteID` - unique identifier
- `StartGrid` - some colonies only have a Startgrid which represents the centre point of the colony/site
- `EndGrid`
- `SampleYear`
- `Common name`
- `AdjustedCount`
- `Unit` - there are several counting units which correspond to the Seabird Monitoring Manual, e.g. AOB = Apparently Occupied Burrow, AON = Apparently Occupied Nest.
- `Accuracy` = counting accuracy entered by recorders. This can be Accurate, Estimate or Estimate hidden an estimate for birds nesting out of view along a cliff for example.
- `Comment` - comment on data.

Questions of interest
The main questions of interest are:

- How to deal best with missing data in the time series and to assess data quality?
- How to describe the trends in seabird abundances?
- To assess which species are showing similar patterns and whether related to spatial location?

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Regression/Flexible Regression (essential).
- Multivariate Analysis (essential).
- Time Series (preferable).

71 Attitudes to Statistics in Introductory Statistics Courses (1)

Project level: Easy/Moderate

71.1 Overall project description

The aim of this project is to examine the relationship between students' performance in introductory university statistics courses and their attitudes towards, and perceptions of, statistics and to compare attitudes/perceptions to statistics in two different introductory statistics courses.

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 more positive attitude would be better able to overcome subject anxiety than other students.

This project will use data obtained from a small study of statistics anxiety that was carried out in the School of Mathematics & Statistics in session 2020-21. Validated instruments measuring various attitudes to statistics were administered to students at the start of two introductory statistics courses: the first ("S1Y") a voluntary course intended for students from a variety of subjects including those who intend to major in maths and/or stats and the second ("BioMed") a compulsory course for specialists in Biomedical Sciences. S1Y had a total of 231 students enrolled and BioMed had a total of 75 students enrolled. Appropriate ethical approval and participant consent were obtained in both courses.

71.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available for each course describing the students' demographics and various attitudes to and perceptions of statistics. Data on the performance of each student in the each course is also included.

The data are stored inthe file *Statistics_Attitudes_Project.csv* and contain the responses to the questionnaire included in the Appendix. In addition to the questionnaire responses, the following columns are included:

- `Course` - the course the student was enrolled on, i.e. "BioMed" or"S1Y"
- `Total Assignments /22` - total mark of continuous assessment (as a mark out of 22)
- `Exam` - final Exam results (out of 100)
- `GPA` - overall Grade Point Average including exam results and all continuous assessment (on 22 point scale)

- `Overall Grade` - overall grade including exam results and all continuous assessment, e.g. “A1”, “B2”, “C3”.

Question(s) of interest
Some questions of interest are:

- What measures of different attitudes to and perceptions of statistics can be derived from the questionnaire responses (e.g. is it possible to average across different responses to get an overall measure of certain attitudes/perceptions?)
- What is the difference between attitudes to and perceptions of statistics between students of different courses, demographic and academic factors and achievement levels?
- What is the relationship of attitudes to and perceptions of statistics with performance, and how do other variables affect these relationships (e.g. demographic and academic backgrounds).

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Statistical Inference.
- Linear Models.
- Data Analysis Skills.
- Generalized linear models.

72 Parameter inference in dynamical systems with nonparametric model (1)

Project level: Moderate/Difficult

72.1 Overall project description

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(KGode) are available to carry out the simulation study, but there is scope for more experienced programmers to code the simulations yourself. More detailed R code examples are availabe in section 3 of this paper:

Niu, Mu, et al. "R package for statistical inference in dynamical systems using kernel based gradient matching: KGode." Computational Statistics 36.1 (2021): 715-747.

72.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data can be generated from the R package KGode by running the code below. The ODE (ordinary differential equation) model used here is the Lotka-Volterra model (prey-predator). From the simulated dataset below, we have y_no, the noised observation of the two states variables in the ODE model. And t_no, the time points for each observation.

```r
require(mvtnorm)
library(KGode)

noise = 0.1  ## set the variance of noise
SEED = 19537
```

```r
set.seed(SEED)

## Define ode function, we use lotka-volterra model in this example.
## we have two ode states x[1], x[2] and four ode parameters alpha, beta,
## gamma and delta.
LV_fun = function(t,x,par_ode){
  alpha=par_ode[1]
  beta=par_ode[2]
  gamma=par_ode[3]
  delta=par_ode[4]
  as.matrix( c( alpha*x[1]-beta*x[2]*x[1] , -gamma*x[2]+delta*x[1]*x[2] ) )
}

## create a ode class object
kkk0 = ode$new(2,fun=LV_fun)
```

```
## ode is sample 2.
```

```r
## set the initial values for each state at time zero.
xinit = as.matrix(c(0.5,1))
## set the time interval for the ode numerical solver.
tinterv = c(0,6)
## solve the ode numerically using alpha=1, beta=1, gamma=4, delta=1.
kkk0$solve_ode(c(1,1,4,1),xinit,tinterv)
## Add noise to the numerical solution of the ode model and
## treat it as the noisy observation.
y_true= t(kkk0$y_ode)
n_o = max( dim( y_true) )
t_no = kkk0$t
y_no = y_true + rmvnorm(n_o,c(0,0),noise*diag(2))
```

Students are also encouraged to try other ODE models such as the SIR model which is widely used in epidemiology.

Questions of interest

The main questions of interest is:

- Simulate data from the Lotka-Volterra model.
- Program the RKHS model and investigate differences in results between the kernels and changing parameters.
- Repeat for the SIR model.
- Are there any differences between the models?

Relevant courses

We recommend that you have taken the following courses to undertake this project:

- Linear Regression Models.

- Kernel Methods.

73 Nowcasting hospital deaths from COVID-19 in England (7)

Project level: Moderate

73.1 Overall project description

Throughout the world, data is collected on infectious diseases to inform planning and action in response to outbreaks. Delayed reporting is where information is not immediately available to decision-makers due to lags in the data collection process. These delays can include time taken to test for diseases, time taken to enter records of new cases and deaths, or time taken for data arriving from local clinics and hospitals to be collated at national surveillance. In a fast moving situation like the COVID-19 pandemic, these delays can muddy the water for decision-makers and leave them forever one step behind while the data catches up.

To address this we can develop statistical models which learn the structure of the reporting delays (e.g. the percentage of cases reported with 1 day on average), and use this learned structure to predict the number of cases or deaths for days where full data isn't yet available.

73.2 Individual project details

How many individual projects are available in this area: 7.

73.3 Project 1 - Nowcasting hospital deaths from COVID-19 in the East of England

Data available
Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for the East of England are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.

- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest
The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in the East of England well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day.* Are you predictions biased and how accurate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

73.4 Project 2 - Nowcasting hospital deaths from COVID-19 in London

Data available
Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for London are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest

The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in London well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day.* Are you predictions biased and how accurate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

73.5 Project 3 - Nowcasting hospital deaths from COVID-19 in the Midlands

Data available
Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for the Midlands are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest
The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in the Midlands well?

- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day*. Are you predictions biased and how accurate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

73.6 Project 4 - Nowcasting hospital deaths from COVID-19 in the North-East and Yorkshire

Data available
Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for the North-East and Yorkshire are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest
The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in the North-East and Yorkshire well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?

- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day*. Are you predictions biased and how accurate are they?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

73.7 Project 5 - Nowcasting hospital deaths from COVID-19 in North-West England

Data available

Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for North-West England are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest

The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in North-West England well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day*. Are you predictions biased and how accurate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

73.8 Project 6 - Nowcasting hospital deaths from COVID-19 in South-East England

Data available
Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for South-East England are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,…
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,…
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest
The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in South-East England well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day.* Are you predictions biased and how accurate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.

- Generalised linear models.
- Linear models.

73.9 Project 7 - Nowcasting hospital deaths from COVID-19 in South-West England

Data available

Data are available on the number of hospital deaths occurring on each day from the 2nd of April 2020 until the 24th of June 2020. Deaths were announced daily by NHS England at 5pm. The death counts are broken down by delay, ranging from 1 to 14 days. Death counts for delay 1 are the number of deaths reported at the first publication date after ocurrence. Counts for delay 2 are the number of deaths reported the following day, and so on. For a given date of death, the total number of deaths is the sum of death counts over all delays. The data are stored in covid_deaths.RData (object name deaths_data) and contain the following columns.

- `region` - Region of England deaths occurred in. Only rows for South-West England are relevant for this project.
- `date` - Date of death.
- `day` - Integer index for the date of death i.e. 1,2,3,...
- `weekday` - Weekday of death, i.e. 1=Monday, 2=Tuesday,...
- `delay` - Delay index for the death counts, as defined above.
- `deaths` - Number of deaths occurring on the given date and reported with the given delay.

Questions of interest

The main questions of interest are:

- Can you design a Generalized Additive Model that fits daily hospital deaths from COVID-19 in South-West England well?
- According to you model, is there compelling evidence that the effect of reporting delays varies with weekday?
- According to your model, how did reporting within the 1st delay increase of decrease over the time period covered by the data?
- Use your model to predict the number of deaths occurring on each day from the 1st of May and ending on the 1st of June. To emulate a realistic nowcasting scenario, for each day you make predictions for *use only data that would have been available on that day.* Are you predictions biased and how accurate are they?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Flexible regression.
- Generalised linear models.
- Linear models.

74 Using Hyperspectral Reflectance Data to define Optical Water Types (1)

Project level: Moderate/Difficult

74.1 Overall project description

Increasingly data on our natural environmental are collected by Earth Observation (EO) satellite technologies with various instruments providing information on our planets physical, chemical and biogeochemical systems. For measurments of water quality, optical instruments are used to obtain data by recording the sun's reflected energy across both visible and infrared bands the wavelength spectra.

It is often, however, not reflectance itself that is the measurement of interest, but instead the determinands such as chlorophyll, total suspended matter and colour dissolved organic matter. To convert reflectance measurements into more interpretable quantities (such as those noted above) a number of 'retrieval algorithms' have been proposed, but such algorithms have been developed based on in-situ measurements from specific water types (e.g. coastal, inland, shallow, turbid etc.) and often perform poorly outwith the water type they are designed for. It is therefore of great importance to be able to identify a what is referred to as an 'optical water type' (OWT) from a reflectance curve to ensure an appropriate retrieval algorithm is applied.

EO satellites, such as the European Space Agency's MERIS (Medium Resolution Imaging Spectrometer) and Sentinel 3 instruments, collect what is known as multispectral data where reflectance is measured at a restricted number of pre-determined bands across the spertrum. Conversely, hyperspectral data, collected using hand-held instruments, collect measurements at a larger number of very narrow wavelength bands, producing an almost continuous measurement every 1nm. Ideally the placement of the specific wavelength bands that are recorded by EO satellites for multivarite measurements are located at wavelengths where there is the most variability across different spectra to maximise our ability to differentiate between different water types.

It is the aim of this project is to explore the variabilty in a large set of hyperspectral data measurements which have been collected from a wide variety of different inland waters (lakes), identifying the wavelengths where there is most differentiation between the spectra. Following this the project will aim to develop a typology of optical water types (OWTs) based on coherence (identifying groups of hyperspectral data which are similar across the wavelength domain).

74.2 Individual project details

How many individual projects are available in this area: 1.

Data available

The data are in the form of 3025 standardized hyperspectral measurements. These data have come from the LIMNADES (Lake Bio-optical Measurements and Matchup Data for Remote Sensing) database which is held and maintained by the University of Stirling and were originally collated from a large number of data sources as part of the Globolakes project (www.Globolakes.ac.uk)

Wavelength is measured in nanometers, denoted by nm, whilst the units of reflectance for this data can be written as 'Standardised R_rs'. Each hyperspectral measurement has an ID label and is comprised of reflectance measurements collected at $1nm$ intervals between wavelenths $400nm$ to $800nm$.

The data are stored in `stlimspec.Rdata` which is a data 3025×401 matrix where the rows correspond to hyperspectral reflectance measurements and columns correspond to the wavelenths (named nm400 to nm800). The rownames give an ID label for each of the spectra.

Question(s) of interest

The main questions of interest are:

- Can we reduce the dimensionality of the hyperspectral curves, capturing the key patterns displayed across the wavelength domain while removing local variability?
- At what wavelengths is most variability in the hyperspectral curves?
- Can we identify coherent groups of hyperspectral curves which display similar patterns of reflectance across the wavelegth domain that could be used to define OWTs?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Multivariate Methods (Essential).
- Functional Data Analysis (Essential).
- Flexible Regression (Desirable).

75 Effect of sibling competition on early childhood growth (1)

Project level: Moderate

75.1 Overall project description

In their article entitled “Sibling Competition & Growth Tradeoffs. Biological vs. Statistical Significance”, available from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150126, Karen L. Kramer, Amanda Veile and Erik Otárola-Castillo describe their research on tracking the growth of children from a Maya community in Mexico to explore the effects of sibling competition on growth during early childhood. In this project, you will use data from 75 children age 2.5-5 to model the children’s growth and explore the extent to which background variables, such as having older or younger siblings, affect patterns of growth.

75.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Data are available on the height and weight of each child at various time points; these are stored in `MayaChildGrowth2007-2011.csv`. The file contains the following 12 pieces of information for every height/weight measurement:

- `KID.ID` the child’s id,
- `MOM.ID` the mother’s id,
- `SEX` the child’s sex,
- `KID.HEIGHT.CM` the child’s height in centimeters,
- `KID.WEIGHT.KG` the child’s weight in kilograms,
- `KID.AGE.YEARS` the child’s age in years (recorded at each measurement time point),
- `MOM.HEIGHT.CM` the mother’s height in 2010,
- `MOM.AGE.YEARS` the mother’s age in 2010,
- `FIELD.SIZE.HA` number of hectares the family has under cultivation, a measure of wealth status,
- `OLDER.SIBLINGS` number of older siblings of the child,
- `YOUNG.SIBS` number of younger siblings of the child,

- FAMILY.SIZE total number of members of the child's family.

Question(s) of interest
The main questions of interest are:

- Fit, and assess the fit of, a simple linear model of height on age, focusing on the age range 2.5 (weaning age) to 5 years. If the linear model is not appropriate, consider a low-order polynomial, for example a quadratic in age.
- Fit, and assess the fit of, a generalised additive model (GAM) to the relationship between height and age.
- Repeat the process to model the child's weight as a function of age.
- Fit mixed models, either linear mixed models or generalised additive mixed models, with a random effect for KID.ID, possibly nested within MOM.ID.
- Add explanatory variables for family size and the number of older and younger siblings to your models. Based on your models, describe the effects (if any) of family size and older/younger siblings on the mean growth pattern for height/weight. Compare your results to those presented in the article "Sibling Competition & Growth Tradeoffs. Biological vs. Statistical Significance".
- Explore the effect of other potential explanatory variables, such as the child's sex, the mother's height and the family's wealth status on the children's mean height and weight growth patterns. Based on your models, describe the effects (if any) of these predictors on the mean growth pattern for height/weight.

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Regression Models.
- Linear Mixed Models.
- Flexible Regression.

76 Patterns in UK regional energy use (1)

Project level: Moderate

76.1 Overall project description

Most of our energy usage comes from electricity and natural gas. The UK Office for National Statistics (ONS) publishes yearly UK regional energy consumption figures for electricity and gas. The aim of this project is to explore the systematic patterns in this data. These include: the change in the usage of electricity and gas in the last 15 years; the comparison of energy usage across private users (houses, largely) and businesses; trying to understand whether any changes are coming from changes in the average amount used per house or business or whether they reflect changes in the number of houses and businesses, and to do all this looking for different patterns in different regions of the UK. Analyses like these form the background to planning for a low-carbon future.

76.2 Individual project details

How many individual projects are available in this area: 1.

Data available
Data are available in comma-separated values (csv) files at https://www.gov.uk/government/statistical-data-sets/stacked-electricity-consumption-statistics-data [Electricity consumption by Government Office Region (GOR) 2005 to 2019] and https://www.gov.uk/government/statistical-data-sets/stacked-gas-consumption-statistics-data [Gas consumption by Government Office Region (GOR), 2005 to 2019].

The structure of the data is described here: https://www.gov.uk/government/statistical-data-sets/stacked-electricity-consumption-statistics-data [Stacked data overview]

Each row is a region of the UK in a particular year. The columns are variables related to electricity or gas consumption, in homes ("domestic") and businesses ("non-domestic"). See **Helpful Starting Hints** for more information.

Questions of interest
Possible questions of interest include:

- How are gas and electricity usage changing in time, both in private homes and in businesses? How does this play into the UK's policy of reduction in carbon emissions?
- Is there any indication that total energy use is changing in time in a different way from average energy use per home/business?
- What is the relationship of the average energy consumption between private households and businesses (gas and electricity taken separately). Is the relationship the same in

different regions? If not, how does it differ and what other factors might be driving that difference?

- What is the relationship of gas and electricity consumption? Is it the same between private households and businesses? If not, how not? Is the relationship dependent on regions of the UK?
- Can you predict regional energy use (and its uncertainty) going forward for a number of years?

Do not feel constrained to answer exactly these questions. You should explore the data enough to come up with some questions of your own to try to answer.

Relevant courses
We recommend that you have taken the following courses to undertake this project:

- Regression modelling.

77 Ordinal classification (1)

Project level: Moderate/Difficult

77.1 Overall project description

Ordinal classification (also called ordinal regression) is a supervised learning problem in-between multinomial classification and regression, whose goal is to predict classes of an ordinal scale such as bad, average, good and excellent. In this project, you will compare standard classification and regression methods with techniques specific to this problem setting. A wide range of specialised approaches can be explored, e.g. methods founded on logistic regression, support vector machine, random forests and Gaussian processes. Some references that could be useful for the project are:

- Gutiérrez, Pedro Antonio, et al. "Ordinal regression methods: survey and experimental study." IEEE Transactions on Knowledge and Data Engineering 28.1 (2015): 127-146.
- Gaudette, Lisa, and Nathalie Japkowicz. "Evaluation methods for ordinal classification." Canadian conference on artificial intelligence. Springer, Berlin, Heidelberg, 2009.

77.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Sixteen benchmark datasets are available from public repositories, including the UCI machine learning repository and the mldata.org repository. The datasets vary in the number of features, examples and classes, with details provided in the table below. For the project, you can choose a few datasets of interest and/or with different data characteristics.

Dataset	#Examples	#Features	#Classes
contact-lenses	24	6	3
pasture	36	25	3
squash-stored	52	51	3
squash-unstored	52	52	3
tae	151	54	3
newthyroid	215	5	3
balance-scale	625	4	3
SWD	1,000	10	4
car	1,728	21	4
bondrate	57	37	5
eucalyptus	736	91	5
LEV	1,000	4	5
automobile	205	71	6
winequality-red	1,599	11	6
ESL	488	4	9
ERA	1,000	4	9

Questions of interest
The main questions of interest are:

- Do advanced methods always provide a better classification performance than standard methods?
- Are some methods more suitable for certain data types than others?
- What measures can be used to evaluate the performance and how appropriate are they?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Data Mining.
- Advanced Predictive Models.

78 Classification with noisy labels (1)

Project level: Moderate/Difficult

78.1 Overall project description

The performance of a classifier heavily depends on the quality of the training data. As labels are manually created, errors are sometimes inevitable and will consequently adversely impact the learned classifier. In this project, you will study one type of label noise of your interest (random label noise, class-dependent label noise or instance-dependent label noise) and investigate one or more types of methods to handle label noise. Methods range from simple remedies, such as training with robust losses and data cleansing using k-nearest neighbours, to advanced techniques, such as Bayesian methods for logistic regression and variants of bagging. A good reference to start is:

- Frénay, Benoît, and Michel Verleysen. "Classification in the presence of label noise: a survey." IEEE transactions on neural networks and learning systems 25.5 (2013): 845-869.

78.2 Individual project details

How many individual projects are available in this area: 1.

Data available

Datasets to be studied will depend on the type of label noises, which are briefly discussed in the reference above. In addition, the following datasets are commonly used and can be analysed in this project: gene arrays in colon cancer data, crowdsourced image datasets, and sentimental text data. Simulated datasets (i.e. simulating label noise) can also be studied.

Questions of interest

The main questions of interest are:

- How sensitive are classical classification methods to label noise?
- How well are the investigated methods in addressing the selected type of label noise? What are their limitations?

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Data Mining.