University of Glasgow

MSc Projects in Statistics 2018/19

Contents

Spatio-temporal prediction of GP prescription rates (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.

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 (main dissertation).
- Spatial statistics or flexible regression (advanced chapter).

Spatio-temporal clustering and trend estimation (2)

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.

Individual project details

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

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

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

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

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.

Individual project details

How many individual projects are available in this area: 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 (main dissertation).
- Spatial statistics (advanced chapter).

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 (main dissertation).
- Spatial statistics or flexible regression (advanced chapter).

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 (main dissertation).
- Time series or flexible regression (advanced chapter).

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

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.

Individual project details

How many individual projects are available in this area: 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 (main dissertation).
- Spatial statistics (advanced chapter).

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 (main dissertation).
- Spatial statistics (advanced chapter).

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 (main dissertation).
- Spatial statistics (advanced chapter).

Identifying contributory factors for disease mortality in England (2)

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.

Individual project details

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

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 (main dissertation).
- Spatial statistics (advanced chapter).

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 (main dissertation).
- Spatial statistics (advanced chapter).

Count data regression models (5)

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.

Individual project details

How many individual projects are available in this area: 5

Number of publications of PhD students

Data available

Data are available regarding the number of publications of 915 PhD biochemistry students during the 1950's and 1960's (on the `Articles` data set which is located in the `Rchoice` library).

A data frame with 915 observations on the following 6 variables.

- *art* Articles during last 3 years of Ph.D.,
- *fem* 1 if female scientist; else 0,
- *mar* 1 if married; else 0,
- *kid5* Number of children 5 or younger,
- *phd* Prestige of Ph.D. department,

- *ment* Articles by mentor during last 3 years.

The first 5 observations can be seen in the table below. Note that the `Articles` data set is already sorted with respect to the number of articles published, hence the 0's at the first column.

articles	gender	marital.status	kids	phd.prestige	mentor
0	male	married	0	2.52	7
0	female	single	0	2.05	6
0	female	single	0	3.75	6
0	male	married	1	1.18	3
0	female	single	0	3.75	26

Note: The original names of these variables are `art`, `fem`, `mar`, `kid5`, `phd`, and `ment`. Two variables (i.e. `fem`, `mar`) have been renamed and transformed to factors.

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

- Is there a relationship between the previous variables and the articles that a Ph.D. student publishes?
- 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 (main dissertation and advanced chapter).

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,

The first 5 observations can be seen in the table below.

children	german	years_school	voc_train	university	religion	year_birth	rural	age_marriage
2	no	8	no	no	Catholic	42	yes	20
3	no	8	no	no	Catholic	55	yes	21
2	no	8	no	no	Catholic	51	yes	24
4	no	8	no	no	Catholic	54	no	26
2	no	8	no	no	Catholic	46	yes	22

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 (main dissertation).

Heavy metal music and negative self-perception

Data available

Download the data set `project_heavy_metal.sav` (attached within the email). Install the package `memisc` and load the data set in R with the command `name_of_data_set <- 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 (main dissertation).

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 (main dissertation).

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.

type	year	period	service	incidents
A	60	60	127	0
A	60	75	63	0
A	65	60	1095	3
A	65	75	1095	4
A	70	60	1512	6
A	70	75	3353	18
A	75	60	0	0

Question(s) of interest

- 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 (main dissertation).

How does the weather affect air pollution levels? (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.

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 (main dissertation).
- Flexible regression/Time Series (advanced chapter).

Investigating free school meals in Scotland (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.

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 (main dissertation).
- Multivariate methods (advanced chapter).
- Spatial statistics (advanced chapter)
- Flexible regression (advanced chapter)

Neighbourhood level crime in Baltimore (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.

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 (main dissertation)
- Spatial statistics (advanced chapter).

Socioeconomic and demographic factors affecting life expectancy (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.

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 (main dissertation)
- Flexible regression (advanced chapter)
- Spatial statistics (advanced chapter).

Lake water Quality (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.

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 (main dissertation)
- Flexible regression (advanced chapter)
- Spatial statistics (advanced chapter).

Modelling obesity in Scotland (5)

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.

Individual project details

How many individual projects are available in this area: 5

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 the BMI distribution 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

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 the BMI distribution 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

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 the BMI distribution 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

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 the BMI distribution 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

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 the BMI distribution 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

Modelling the progression of world records in athletics (3)

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.

Individual project details

How many individual projects are available in this area: 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.

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.

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.

Modelling the growth of infants in Glasgow (3)

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.

Individual project details

How many individual projects are available in this area: 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

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

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

Identifying Seasonal Coherence in Global Lake Surface Water Temperature (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.

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 and Linear Models (main dissertation).
- Functional Data Analysis or Flexible Regression (advanced chapter).

Temporal Patterns in Temperature and NDVI (2)

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.

Individual project details

How many individual projects are available in this area: 2

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 or Time Series (main dissertation).
- Environmental statistics or flexible regression (advanced chapter).

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 or Time Series (main dissertation).
- Environmental statistics or flexible regression (advanced chapter).

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

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.

Individual project details

How many individual projects are available in this area: 6

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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- flexible regression and environmental statistics.

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`, `TPSl8.csv`, `TPSD6.csv`, `chlaRB5.csv`, `chlaSl8.csv`, `chlaSD6.csv` and `sdepthRB5`, `sdepthSl8.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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- flexible regression, environmental statistics and linear mixed models.

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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- flexible regression and environmental statistics.

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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- linear mixed models, flexible regression and environmental statistics.

Temporal patterns and drivers of water quality at Bassetnhwaite 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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- flexible regression and environmental statistics.

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 and time series (main dissertation).

The following courses/topics may be useful for the advanced chapter:

- flexible regression and environmental statistics.

Modelling housing price substitutability in Glasgow (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 substitutability, which is the idea that items with different characteristics may be exchangeable to a consumer. In the case of the housing market, this may mean that a dwelling of a certain type in one area or intermediate geography may be substituted for a similar dwelling in another intermediate geography entirely. The questions are, how can we visualise this substitutability data (which is pairwise rather than individual observations) and can we group together substitutable areas to form housing sub-markets?

This project will look at constructing a substitutability matrix with a measure of substitutability between all pairs of intermediate geographies in the Glasgow City Local Authority region. This can then be used to visualise how the intermediate geographies are substitutable in terms of housing. The methods used to explore these will include multidimensional scaling and cluster analysis.

Individual project details

How many individual projects are available in this area: 1

Data available

Data are available on median yearly house prices air 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:

- What does the distribution of IZ's look like in terms of substitutability?
- What are the clusters that make up the housing sub-markets in terms of substitutability?

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

- Multivariate Methods (main dissertation).
- Spatial statistics (advanced chapter).

Modelling housing price change in Glasgow (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 pattern of change over time in housing areas in Glasgow. The questions are, how can we visualise this change and can we group together areas that have similar change over time to form housing sub-markets?

This project will look at cluster analysis of yearly media house price time series for intermediate geographies in the Glasgow City Local Authority region. This can then be used to visualise how the intermediate geographies form submarkets. The methods used to explore these will be cluster analysis.

Individual project details

How many individual projects are available in this area: 1

Data available

Data are available on median yearly house prices air 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:

- What does the pattern of house price change in IZ's look like?
- What are the clusters that make up the housing sub-markets in terms of house price change?
- Can we model the time series as smooth curves and cluster these?

Relevant courses

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

- Multivariate Methods (main dissertation).
- Spatial statistics (advanced chapter).

Cluster Analysis of Acute Respiratory Distress Syndrome (2)

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.

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

Individual project details

How many individual projects are available in this area: 2

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 (main dissertation).

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 (main dissertation).

Classification Analysis of Acute Respiratory Distress Syndrome (2)

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.

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

Individual project details

How many individual projects are available in this area: 2

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 (main dissertation).

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 (main dissertation).

Bayesian way of predicting the price of gold (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.

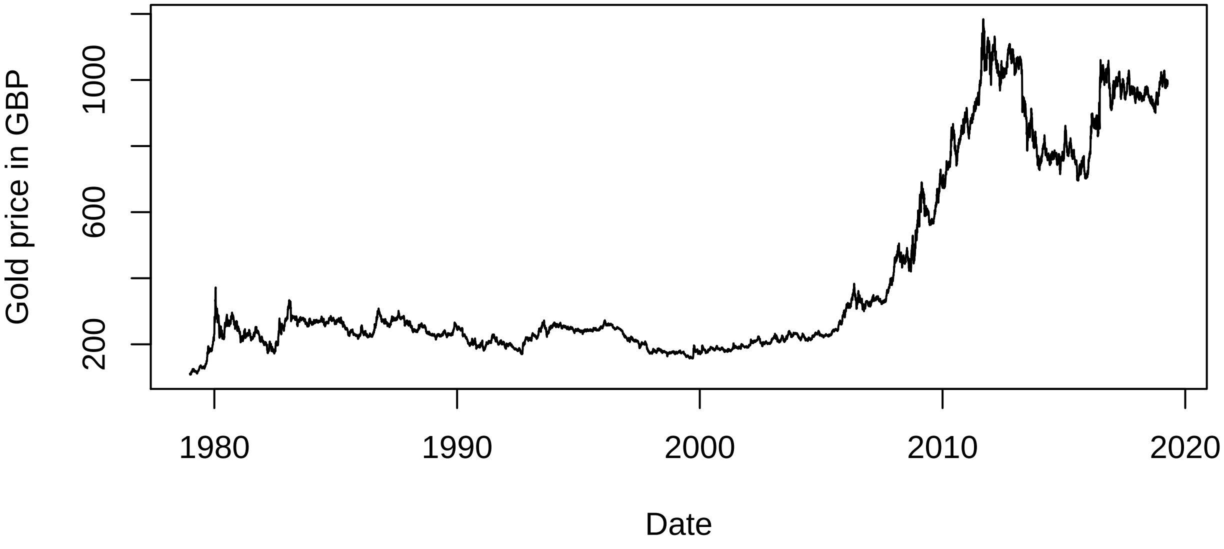
Such an analysis requires inferring model parameters that match current history of the price. Unfortunately, the likelihood imposed by the SDE models does not have a closed form, and therefore traditional inference methods are not applicable to this problem.

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

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:

- How to formulate informative priors for a problem with good intuitive understanding of the subject?
- How to perform inference of model parameters when working with SDE?
- How to make predictions 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.

Bayesian Linear Models and Bayesian Lasso (4)

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.

Individual project details

How many individual projects are available in this area: 4

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:

- 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 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.

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.

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.

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:

```r
require(mlbench)
```

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

```r
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.

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

Overall project description

Please notice that the availability of these projects is dependent on ethic approval.

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.

Individual project details

How many individual projects are available in this area: 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

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

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

Identifying risky bank loans (1)

Overall project description

The global financial crisis of 2007-2008 has highlighted the importance of transparency and rigour in banking practices. As the availability of credit has been limited, banks have tightened their lending systems and are turning to statistical decision support systems to more accurately identify risky loans. Such automated credit scoring models can be deployed for instantly approving credit applications on the telephone and the web. The objective of the present project is to predict whether a loan went into default based on a set of various explanatory attributes. These attributes include indicators for credit history, purpose of loan, credit amount, employment status, gender, age, personal status, housing, and the applicant's type of profession. The task of this project is to develop and assess a classifier that reads in these attributes and correctly predicts whether the loan is going into default.

Individual project details

How many individual projects are available in this area: 1

Data available

The data are available in file `creditData.txt`. The first line is a standard line of headings, where the first column (y) indicates whether or not the loan went into default (0:no, 1:yes), and the following 24 columns (x01 to x24) are 24 real numbers with the attributes mentioned above.

Question(s) of interest

How accurately can the risk of loan default be predicted with a statistical classifier? How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant courses

Inference, Flexible regression, Generalized linear models, Mutivariate methods, Big data analytics, Introduction to R programming

Optical character recognition (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 recognize 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:

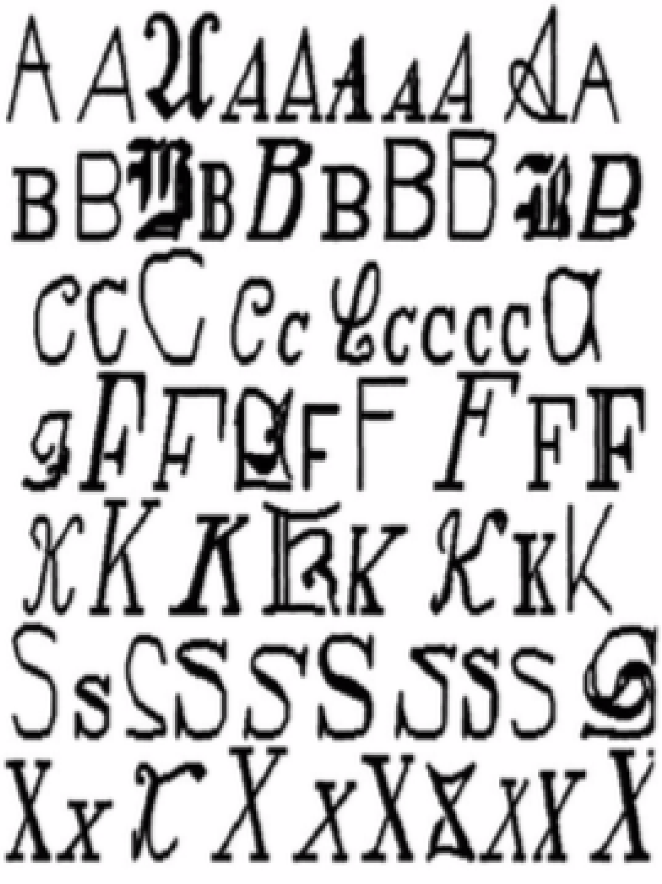


Figure 1. Examples of the character images generated by "warping" parameters.

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.

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?

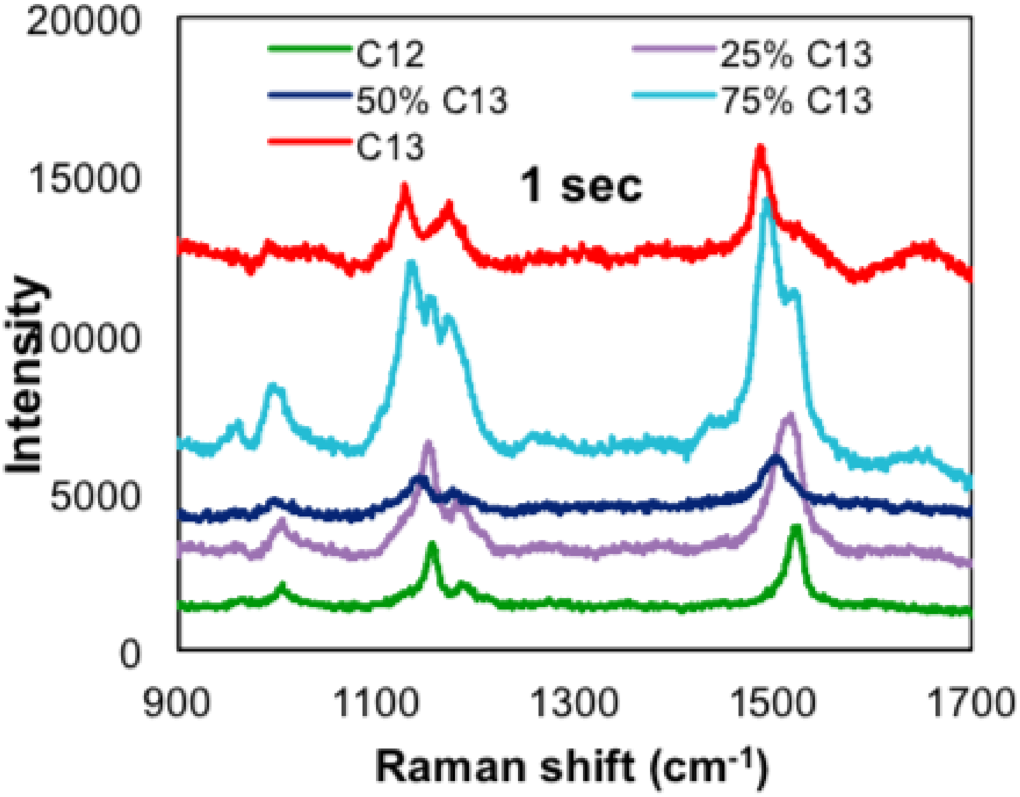
Relevant courses

Inference, Flexible regression, Generalized linear models, Mutivariate methods, Big data analytics, Introduction to R programming

Classifying bacterial metabolic states with Raman spectroscopy (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.

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? How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant courses

Inference, Flexible regression, Generalized linear models, Multivariate methods, Big data analytics, Introduction to R programming

Reading attainment in primary school children (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)

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

Finding epigenetic signatures for human ageing (2)

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.

Individual project details

How many individual projects are available in this area: 2

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 or Machine Learning (main dissertation).
- Multivariate methods or Machine Learning (advanced chapter).

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 or Machine Learning (main dissertation).
- Machine Learning or Bayesian Statistics (advanced chapter).

Determining genetic variation associated with heart disease (2)

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.

Individual project details

How many individual projects are available in this area: 2

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 or Machine Learning (main dissertation).
- Big Data Analytics, or Bayesian Statistics/Advanced Bayesian methods (advanced chapter).

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 or Machine Learning (main dissertation).

- Big Data Analytics, or Bayesian Statistics/Advanced Bayesian Methods (advanced chapter).

Statistical analysis of US presidential election data (2)

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.

Individual project details

How many individual projects are available in this area: 2

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 or Machine Learning (main dissertation).
- Big Data Analytics, Multivariate methods or Machine Learning (advanced chapter).

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 or Machine Learning (main dissertation).
- Big Data Analytics, Machine Learning or Bayesian Statistics (advanced chapter).

Predicting Olympic medal counts (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.

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.

```r
oldat <- read.csv("olympics2016.csv")
oldat$gdp16 <- as.numeric(oldat$gdp16)
head(oldat,1)
```

```
##       country country.code gdp00 gdp04 gdp08 gdp12 gdp16 pop00 pop04 pop08
## 1 Afghanistan          AFG  #N/A  5285 10191 20537    33 20094 24119 27294
##   pop12 pop16 soviet comm muslim oneparty gold00 gold04 gold08 gold12
## 1 30697 34656      0    0      2        0      0      0      0      0
##   gold16 tot00 tot04 tot08 tot12 tot16 totgold00 totgold04 totgold08
## 1      0     0     0     1     1     0       298       301       301
##   totgold12 totgold16 totmedals00 totmedals04 totmedals08 totmedals12
## 1       301       298         915         924         949         956
##   totmedals16  bmi altitude athletes00 athletes04 athletes08 athletes12
## 1         949 23.3     1790          0          5          4          6
##   athletes16 host
## 1          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 (main dissertation)
- Generalised Linear Models (main dissertation and advanced chapter)

Linear Mixed Models and Flexible Regression may also be useful for the advanced chapter.

Monitoring physical and mental health outcomes in Scotland (2)

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.

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

Individual project details

How many individual projects are available in this area: 2

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 may also be useful for the extension chapter.

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 may also be useful for the extension chapter.

Self-rated health and socioeconomic status in Scotland (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.

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 may also be useful for the extension chapter.

Using the urinary steroid profile to detect prostate cancer in men (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.

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 (main dissertation and advanced chapter)
- Generalised Linear Models (main dissertation and advanced chapter)

Does playing Pokemon Go increase physical activity? (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.

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.

```r
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
## 1     Strongly agree     Strongly agree           Disagree
##   attitude_attitude4 ATTENTION_filter1 attitude_attitude5
## 1     Strongly agree          Disagree     Strongly agree
##   attitude_attitude6 stepsattitude_attitudeB1 stepsattitude_attitudeB2
## 1     Strongly agree                        7                        1
##   stepsattitude_attitudeB3 stepsattitude_attitudeB4
## 1                        7                        5
##   stepsattitude_attitudeB5 stepsattitude_attitudeB6
## 1                        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 (main dissertation and advanced chapter)

Generalised Linear Models may also be useful.

Detecting magnetic storms in space (4)

Overall project description

Space weather refers to electromagnetic disturbances in the near-Earth environment as a result of the Sun-Earth interaction. Severe space weather events such as magnetic storms can cause disruption to a wide range of technologies and infrastructure, including communications systems, electronic circuits and power grids. Because of its high potential impact, space weather has been included in the UK National Risk Register since 2011.

Space weather monitoring and early magnetic storm detection can be used to mitigate risk in sensitive technological systems. The aim of this project is to investigate the electromagnetic disturbances in the near-Earth environment through developing statistical models that quantifies the variations and uncertainties in the near-Earth magnetic field. In addition, information on the presence of magnetic storms in the near earth environment have been compiled independently. It is of interest therefore to examine how the signal in the satellite measurements may show early warning of impending storms

Individual project details

How many individual projects are available in this area: 4

Data available

Projects 1-4 are based on the 4 different satellites

Data of the near-Earth magnetic field arise from satellites. The Cluster II mission (Escoubet et al., 2001a) has four satellites that provide in-situ measurements of the near-Earth magnetic field at time-varying locations along their trajectories. The Cluster II mission has four satellites moving in a tetrahedron shape and sampling the magnetic field as they orbit around the Earth. The four datasets comprise the orbital average of each Cluster II satellite measurements of the near-Earth magnetic field. The magnetic field vectors have three components and are measured as Bx, By and Bz in the unit of nano-Tesla (nT). Each data file record measurements from one satellite over 11 years from 2003-2013. The data are stored in `sat_xxx.csv` where `xxx` is 1-4. In addition, from other measurements, there are independent magnetic storm information, specifically periods when a magnetic storm occurred and its duration available separately in `storm.csv`.

There are 6 columns in each satellite dataset. `NA` represents a missing value;

- `year`
- `month`

- `orbit`
- `Bx`
- `By`
- `Bz`

Question(s) of interest

The main questions of interest are:

- How does the magnetic field components, Bx, By, and Bz, vary in time and are the variations related?
- Is there a change in the magnetic measurements as a precursor to a storm condition (evidence of a changepoint either in the mean or variance of the magnetic field)?.

Relevant courses

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

- Time series (main dissertation)
- Flexible regression (advanced chapter)

Modelling isotope composition over space (1)

Overall project description

Migration in insects is widespread. Linking locations used by individuals is crucial to understand their ecology, to conserve threatened species and to understand how climate change may change their numbers and distribution. We can study the environment that the insects live in through studying isotopes. Isoscapes, based on the GNIP hydrogen isotope database may help. The GNIP data support the analysis of the temporal and spatial variations of environmental isotopes in precipitation (mainly oxygen-18 and deuterium (hydrogen)) and provides basic data for the use of isotopes in hydrological investigations but is limited in the UK since it involves only single measuring locations. Therefore this project is focussed on an extensive study of 4 different isotopes which were measured across the UK in many locations and based on a resident insect. The insect chosen is the Brimstone Moth.

Individual project details

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

Data available

299 moths were collected from 93 locations, and the time of year when collected was recorded by the volunteers. The 4 elements of interest are hydrogen, sulphur, carbon and nitrogen. These have been measured and reported in a variety of ways, including the isotope ratio (cols 7-10), the weight of the carbon, nitrogen, hydrogen and sulphur in the sample (cols 11-13) and finally the % weight of each element (cols 14-16).

Data available for each moth sample are:

- `Location` - latitude and longitude
- `Date collected` - in dd-mmm-yy format
- `Weight of sample` - in mg
- `Temperature` - note there are missing values
- `d2H-DS` - hyrdrogen isotope ratio
- `d15N lin` - nitrogren isotope ratio
- `d13C lin` - carbon isotope ratio
- `N mg, C mg, S mg` - Nitrogen (N), Carbon (C) and Sulphur (S) weight in mg
- `N %, C %, S %` - Nitrogen (N), Carbon (C) and Sulphur (S) weight as a %
- `d15N mean, d15N sd, d13C mean, d13C sd, d34S mean, d34S sd` - Nitrogen, sulphur, carbon mean isotope ratio for each location and the associated standard deviation.

Question(s) of interest

The main questions of interest are:

- To explore the spatial distribution in the 4 (carbon, hydrogen, sulphur and nitrogen) isotopes in a resident insect, and to determine whether there are well defined isotopic gradients which can be explained in terms of climate or other locational variables.

Relevant courses

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

- Linear models (main dissertation)
- Spatial methods (main dissertation)
- Flexible regression (advanced chapter)

Modelling wind speed at a windfarm site (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

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 (main dissertation)
- Flexible regression (advanced chapter) (quantile regression and extremes).

Exploring a household's energy consumption (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?

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

How well can you establish the geographical origin of a DNA sequence? (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.

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.

Supervised statistical classification (2)

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.

Individual project details

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

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:

- Generalized linear models.
- Multivariate methods.

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:

- Generalized linear models.
- Multivariate methods.

Spatio-temporal modelling of big environmental data (3)

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.

Individual project details

How many individual projects are available in this area: 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 (advanced chapter).

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 (advanced chapter).

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 (advanced chapter).

Modelling sunspot numbers (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.

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. All are available there.

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

Investigating properties of the G test for Hardy–Weinberg equilibrium by simulation (1)

Overall project description

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.

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

Exploring variation in human skull shape (2)

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.

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.

Relevant courses

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

- Statistical Inference, Regression Modelling, GLMs, Multivariate methods.

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?

Classification approaches to the cranial data

Questions of interest

- 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?

Usage of hire bikes in London (1)

Overall project description

As part of its open data framework Transport for London (TfL) release 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 analyise 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.

Individual project details

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

Data available

The data for this project needs to be downloaded from the TfL website. Cycle usage data can be downloaded from https://cycling.data.tfl.gov.uk/. A sample dataset is also given to students at the start of the project.

Question(s) of interest

Which bike stations are the most popular? What types of trips are the bikes used for? What are the detailed spatio-temporal patterns? What time of the day is most popular? Are weekdays different from week-ends? Is summer different from winter?

Relevant courses

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

- Regression Models
- Generalised Linear Models
- Time Series and/or Flexible Regression might be helpful

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

Overall project description

On May 2nd, local elections were held in England. The Conservatives fared very badly, but Labour didn't 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.

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 interests 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 might be helpful

Predicting astrophysical properties of stars based on their light spectrum (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.

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 or Machine Learning would be helpful

The @realDonaldTrump or not? (1)

Overall project description

The tweets posted by @realDonaldTrump are either written by himself or his team. Tweets from Donald Trump himself tend to be angier and more negative than the ones sent his his staff.

Donald Trump is known to have used a Samsung Galaxy smartphone to write his tweets, whereas his team uses iOS devices. The device used to send a tweet can be retrieved using the Twitter API, so it used to be easy to determine whether a tweet was written by Donald Trump himself or his team. However, in March 2017, Donald Trump switched to an iPhone, so these days there is no way of telling whether Donald Trump has written a tweet himself.

In this project you build a model that will learn from Donald Trump's tweets from up to March 2017 what characterises a real Trump tweet and use this to predict for

Many people have tried this, see for example @ReallyIsTrump, https://www.evolvedatascience.com/post/reallyistrump-tweet-predictor/ or http://didtrumptweetit.com/

Your task in this project is to develop statistical models to improve upon these.

Donald Trump's role in the project is limited to providing tweets. I'm afraid he won't be available to come along to any meetings (who would want that anyway . . .).

Individual project details

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

Data available
The data made available consists of the text of the tweet, the day of the week it was sent, the decimal hour when it was sent and an indicator whether it was sent by Trump (based on the device used to send it).

Question(s) of interest
The main question of interest is to predict whether the tweet sent by Trump or not. The only two covariates provided are the day of the week and the decimal hour. You need to generate other covariates from the text of the week.

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

- Generalised Linear Models

Housing market analysis (3)

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

Individual project details

How many individual projects are available in this area: 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 (main dissertation).
- Generalized Linear Models, Advanced Data Analysis, Big data Analysis (advanced chapter).

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 (main dissertation).
- Advanced Data Analysis (advanced chapter).

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 (main dissertation).
- Advanced Data Analysis (advanced chapter).

Grocery sales data analysis (3)

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.

Individual project details

How many individual projects are available in this area: 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 (main dissertation).
- Generalized Linear Models, Advanced Data Analysis, Big data Analysis (advanced chapter).

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 (main dissertation).
- Advanced Data Analysis (advanced chapter).

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 (main dissertation).
- Advanced Data Analysis (advanced chapter).