Exploring the effect of weather and climate on official statistics: Guidance for official statistics producers

Version: 2.0
Date: November 2016
Author: ONS Time Series Analysis Branch
Date of next review: November 2018

Does the weather or climate affect your data?

If you want to analyse your data to see if they are affected by weather or climate, you will need to know:

1. What weather or climate data are available.
2. Appropriate methods and tools for analysis.
3. How to communicate results to users.

This guidance is designed to assist producers of official statistics with accessing and using weather and climate data, conducting an analysis and communicating results to users.

Who is this guidance for?

Producers and analysts of official statistics interested in exploring the use of weather or climate data to help them interpret their own data. Concepts discussed in some sections can become technical and some familiarity with concepts such as regression is assumed.

How should this guidance be used?

The interactive flow chart on the next page provides an overview of the layout of the guidance. While there is an order to the guide it is not written with the intention that users will read it cover to cover.

- **Want to find out what weather and climate data are available?** Take a look at the Met Office guide to weather and climate data in section 2.2
- **Interested in seeing examples of weather analysis and how it was carried out?** The examples section of the guide includes examples of analysis carried out on retail sales, road accidents or ambulance repsonse times. Each example includes data and code so that the analysis can be replicated as a learning tool.
- **Starting to plan your own analysis?** The flow chart on the next page provides an overview of how to conduct an analysis with links taking you to relevant sections of the guidance.

1. Plan your analysis

1.1. Clarify objectives

1.2. Methods for analysis

2. Obtaining weather and climate data

2.1. Selecting data

2.2. Met Office guide to weather and climate data

3. Initial data analysis

3.1. Assessing data quality and structure

3.2. Further preliminary analysis

4. Time series modelling

4.1. Time series concepts

4.2. Time series models

4.3. Software

5. Communication

5.1. Interpreting your results

5.2. Presenting results to users

EXAMPLES

6. Retail Sales

Overview

Analysis

Data and code

7. Road accidents

Overview

Analysis

Data and code

8. Ambulance response times

Overview

Analysis

Data and code

9. FAQs

Contacts

Methodology Advisory Service

Met Office

Time Series Analysis Branch

Good Practice Team

External links

Met Office

Methodology Advisory Service

X-13ARIMA-SEATS

1 Plan your analysis

The focus of this guidance is on planning a time series investigation of your data to help you determine potential weather or climate effects on your time series. Of course you are not restricted to time series analysis when considering the affect of weather or climate on official statistics and many of the points are relevant to other types of statistical investigation. Planning your analysis will involve planning each of the steps described in this guidance. It is important to consider how the results will be used and communicated to users and so should be an integral part of the planning process. As with any project planning, this should be an iterative and responsive process whereby plans are revisited and modified as appropriate.

It is always advisable to conduct a thorough literature review and adapt your planned investigation as necessary. This will be useful for comparisons and evaluations of results and for providing informed communication of the analysis.

Back to flow chart

1.1 Clarify objectives

Before attempting to construct a model, it is important to formulate a clear research question and clarify what the investigation should achieve. Formulating the question will help in deciding data requirements, appropriate approaches to modelling and working out whether the data and methods of analysis will address the question posed.

The requirements for the analysis may have resulted from a question about movements in a particular time series or from some recent notable weather or climate related event. For example, if there has been particularly heavy snow, storms or floods, then users of the data may ask if a series has been affected by those events. This may lead to a more general question, such as, how is this time series affected by the weather and climate?

This is quite a general research question and to ensure that users will be satisfied with the outcome of any analysis it is important to consider more precise questions, such as, are my data affected by deviations from the average level of rainfall in a month? At this stage it can be helpful to think about what sort of answers might be presented. When the objectives are clear it will become more straightforward to plan the investigation and formulate hypotheses to test in the modelling.

Back to flow chart

1.2 Methods for analysis

The methods for analysis considered in this guidance include simple graphical analysis and time series modelling. However, other methods could also be considered. The following methods have been found in some of the literature on assessing the impact of weather and climate on official statistics.

- **Visual analysis**

  Plot the time series being analysed and a weather variable on the same chart and look for any visual relationships. This has been used within the GSS in an ONS Retail Sales article, entitled How sensitive to the weather is the retail sector? (ONS, 2014). The seasonally adjusted Retail Sales Index was plotted on a graph with deviations from the average temperature. This led to identifying peaks and dips in the index that coincided with higher or lower than average temperature.

- **Ordinary least squares regression**

  Simple ordinary least squares regression has been used widely in literature, one example being (Smith, 1982) where numbers of accidents on wet roads was modelled against monthly precipitation. The Department for Transport (DfT) have also looked at modelling the irregular component of a decomposed time series against weather effects following on from a Masters dissertation which also explored weather. Care should be taken with this approach, especially when using time series as it assumes observations are independent, and time series often do not satisfy this assumption.

- **Matched pairs analysis**

  A paper analysing the effect of weather conditions on road accidents in Glasgow used matched pair type analysis to investigate effects (Smith, 1982). Daily numbers of road accidents, along with daily weather data on whether conditions were wet or dry were used to produce matched pairs of days one week apart, where one day was wet and the other dry. The average numbers of accidents on wet and dry days were compared and t-tests were used to test for any significant differences.

- **regARIMA modelling**

  A discussion paper published by Statistics Netherlands used regARIMA modelling with weather effects included as regressors to analyse GDP data (CBS, 2014). This technique is also used by the Department of Energy and Climate Change (DECC), whom produce temperature adjusted energy consumption statistics, which are adjusted using factors calculated by fitting a regARIMA model with weather regressors (Rahman, 2011).

Back to flow chart

2 Obtaining weather and climate data

2.1 Selecting data

Users of this guide are likely to have a time series already in mind that they would like to investigate further. For the purposes of constructing a time series model it is important to consider what type of weather and climate data are appropriate for your model building. Below are some questions that should be considered in relation to your own data and also in relation to weather and climate data. Hover over the questions for a more detailed explanation. The full explanations are available in Annex A.

First, consider aspects of your data.

- What do your data measure?
- Do you have flow or stock data?
- What is the span of your data (start and end date)?
- What is the geographic coverage of your data?
- What is the frequency of your data?
- What types of weather or climate variables might be expected to affect your data?
- How might your data be affected by the weather and climate?
- Are there any details of the compilation process of your data that need to be considered in your analysis and in your selection of appropriate weather variables?
- How are your data currently published, and how might your publications be affected by the analysis?

Next, consult the section 2.2 of this guide on weather and climate statistics to find out what sort of data are available. For the purposes of starting to build a time series model you will need to consider the aspects of your data outlined above and aspects of the weather or climate data:

- Does the geographic coverage of the weather and climate data match that of your data?
- Would a bespoke weather or climate data time series weighted by some other variables be appropriate?
- Does the frequency of the weather and climate data that you are interested in using in your analysis match that of your data?
- How should the weather or climate data be aggregated over time or geography?

- Are there alternative derivations of weather and climate variables from the available weather data that could be appropriate for your analysis?
- Does the span match that of your data?
- What are the relationships (correlations) between the weather and climate variables of interest?

Back to flow chart

2.2 Met Office guide to weather and climate data

Met Office produce a range of weather and climate data products. Some of the general data types are described in table 1, with more detailed description in the rest of this section.

Back to flow chart

Table 1: *Overview of available weather data types.*

Type	Description	Recommended use	Example application
Climate summaries	Daily, monthly, seasonal and annual weather and climate summaries.	Placing weather and climate events into a descriptive context.	Met Office Climate Summaries
Climate series	Homogeneous national and regional climate indices derived from station observations.	Monitoring UK climate variability and change over time. Provides a peer-reviewed method of summarising weather and climate observations at a national and regional scale.	Retail Sales Road Accidents
Gridded climate data	Daily, monthly, and long term average observations interpolated onto a uniform grid.	Monitoring spatial patterns of UK climate variability and change. Provides a peer reviewed method of managing change in observing network (sampling error) through time.	Climate of the UK and recent trends
Station climate statistics	Monthly and long term average climate statistics determined for long-running UK climate stations.	Monitoring climate variability and change at specific locations.	Ambulance response times
Daily climate observations	Archive of daily weather and climate observations.	Daily observations of a range of meteorological elements at specific locations.	Daily temperature used to weather-adjust energy consumption figures published by DECC
Hourly observations	A range of observations are made at hourly resolution, primarily from the Met Office network of automatic weather stations.	Hourly observations for applications that require resolving the diurnal cycle or sub-daily weather.	Met Office Datapoint

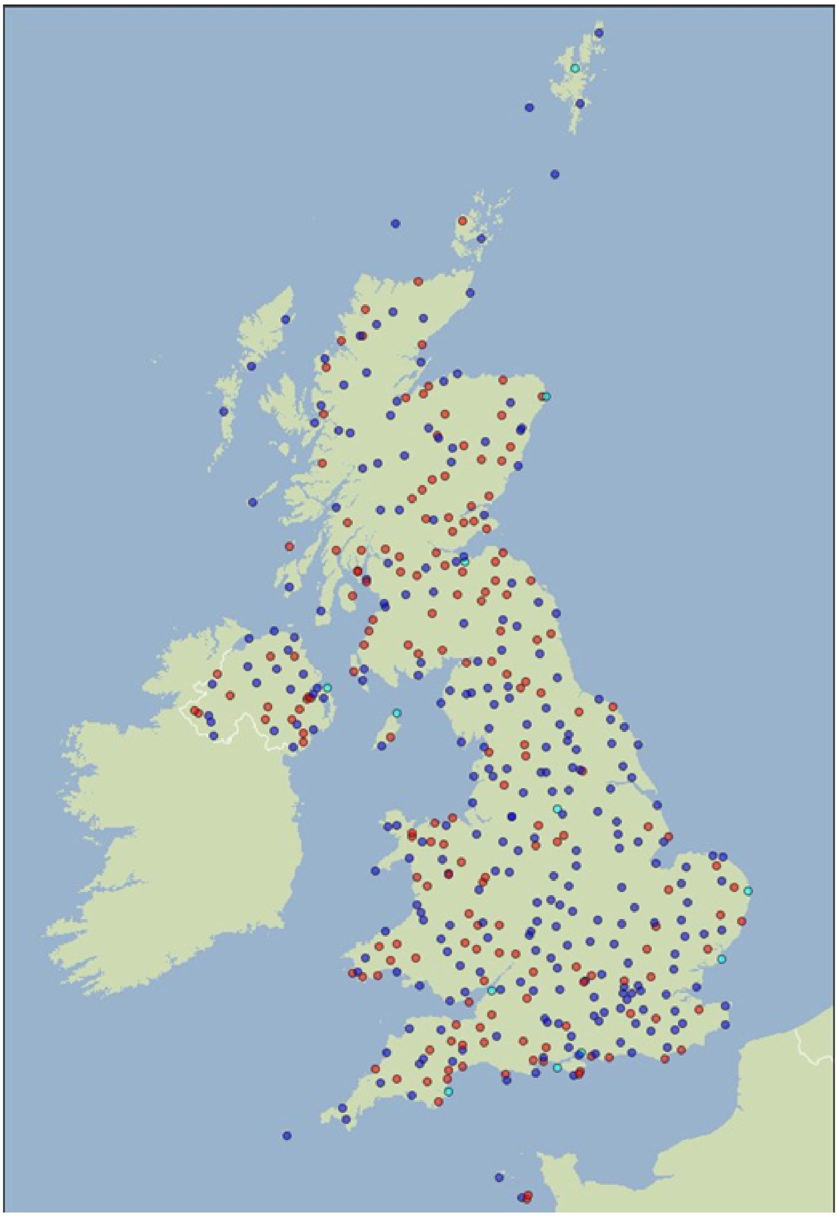
2.2.1 Observing weather

The observations programme at the Met Office support a wide variety of observing systems for the UK from in-situ land observations, upper air measurements from weather balloons, marine observations from ships and buoys, and remote sensing from radar and satellites. More information about weather data from the Met Office can be found here. Here we concentrate on weather observations over land which reflect the weather as most people experience it.

Observations that are made to provide information on the present state of the atmosphere and to serve the requirements of operational numerical weather prediction are termed synoptic observations. Observations from around 270 UK synoptic stations are collected in real time with a spacing such that typical frontal weather systems can be detected by the network. However some smaller features, for example localised thunderstorms could evade the surface network entirely. The network is supplemented by additional observations such as 170 co-operating climate stations providing additional observations to support climate monitoring or other specific applications, or a network of several thousand rain gauges managed the Environment Agency for hydro-meteorological monitoring. For instance, hourly temperature is a synoptic element, but daily maximum and minimum temperature are climate elements; cloud cover is a synoptic element but hours of bright sunshine over the day is a climate element.

Automation of the meteorological observing system (blue stations in figure 1) means that much of the network now reports data every minute to the Met Office headquarters in Exeter. The volume of minute resolution data is too much for most applications to practically handle so the data are encoded to international standards with 24 hourly synoptic observations per day, and up to 3 climate observations. The climate observations are typically made at 0900 UTC, with some sites making a further observation at 2100 UTC (for example maximum daytime temperature between 0900 and 2100). Observations from manual climate stations may come in as collectives of data more sporadically so for comprehensive climate assessments.

Figure 1: *The current meteorological observing network showing automated (blue) and manual (red) observing sites*



All climate stations record daily maximum and minimum air temperature and rainfall amount, recorded over the period 0900-0900 UTC period (i.e. 1000-1000 during the summer). Many observe additional elements and the principal elements for climate applications include:

- Maximum air temperature at 1.25 m above the ground (0900 UTC to 0900 UTC the next day)
- Minimum air temperature at 1.25 m above the ground (0900 UTC to 0900 UTC the next day)
- Air temperature at 1.25 m above the ground (0900 UTC)
- Grass minimum temperature (dusk to 0900 UTC the next day)
- Soil temperature at 0.1 m, 0.3 m and 1.0 m (0900 UTC)
- Relative humidity at 1.25 m above the ground or the wet bulb equivalent (0900 UTC)
- Rainfall amount (0900 UTC to 0900 UTC the next day)
- Depth of lying snow (0900 UTC)
- Sunshine duration (0000 UTC to 2300 UTC)

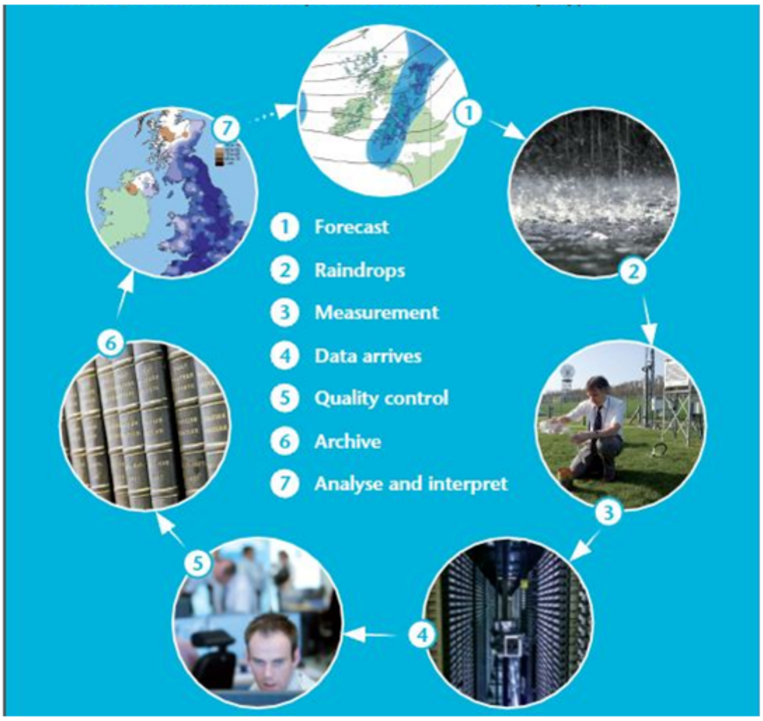
The current network of synoptic and climate stations covering the UK as shown above can be found here.

A factsheet with more detail on meteorological observations can be found here.

Back to flow chart

Back to Data Summary table

2.2.2 Handling weather and climate data



Meteorological observations go through a number of processing and quality assurance steps from measurement to inclusion in the climatological data archives. In order to promote best practice, and avoid duplication of effort the Met Office produce a range of weather and climate data products designed to provide consistent and peer reviewed methodology for handling weather and climate data. In particular those aspects relating to reporting practices, sampling, representativity and homogeneity. Therefore it is advised that users review their requirements carefully alongside the range of existing data types.

Things to consider:

- Reporting practice
- Diurnal and seasonal cycle
- Sampling - representativity
- Trends and autocorrelation
- Dependent variables

2.2.2.1 Reporting practice

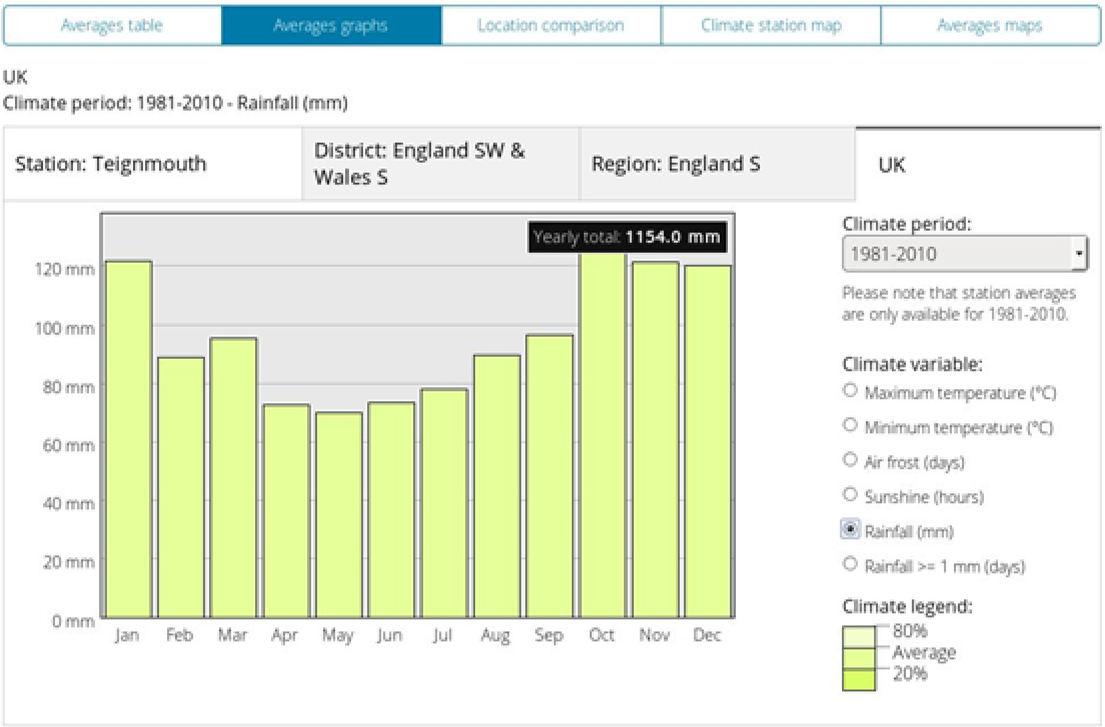
It is important to consider how the reporting practices of different meteorological observations might influence your statistical analysis. For example minimum temperatures (Tmin) are normally reached in the early morning before sunrise followed by heating through the day to peak at maximum temperatures (Tmax) in the early afternoon. We have not always had the technology to measure the true mean temperature (Tmean) over a 24 hour period, however maximum and minimum thermometers have existed for several centuries so for consistency standard climatological practice is to estimate Tmean as the average of the reported Tmax and Tmin. Observational practice is to report readings of Tmax and Tmin over a designated time period, usually 12 or 24 hours. For a 24 hour observation made at 09:00 GMT this means that the associated Tmin is most likely to have occurred on the same calendar day as the report, and the Tmax to have occurred the previous afternoon. Therefore it is standard practice to throw back the Tmax observation so that it is associated with the day in which it was most likely to have occurred. However it should be noted that a number of meteorological conditions, for example the passage of a warm front during a winter night or a foehn wind in the lee of mountains, could result in marked departures from a normal diurnal cycle. Similarly daily rainfall accumulation is normally reported over a 24 hour period 09:00 to 09:00. Therefore the rainfall is most likely but not entirely associated with the preceding day.

In summary even for the most ubiquitous of meteorological parameters, temperature and rainfall, there are a number of potential considerations for the treatment of the data relating to how the observations and made, and what level of subsequent processing has been carried out on them, and how they relate to the statistics of interest. Met Office archives will store data as they were reported, but derived climatological products such as the Gridded Climate Data and Climate Series will have accounted for the factors described above in a consistent manner following climatological best practice.

Back to flow chart

Back to Data Summary table

2.2.2.2 Diurnal and seasonal cycle



UK
Climate period: 1981-2010 - Rainfall (mm)

Station: Teignmouth | District: England SW & Wales S | Region: England S | UK

Climate period:
1981-2010

Please note that station averages are only available for 1981-2010.

Climate variable:
○ Maximum temperature (°C)
○ Minimum temperature (°C)
○ Air frost (days)
○ Sunshine (hours)
● Rainfall (mm)
○ Rainfall >= 1 mm (days)

Climate legend:
80%
Average
20%

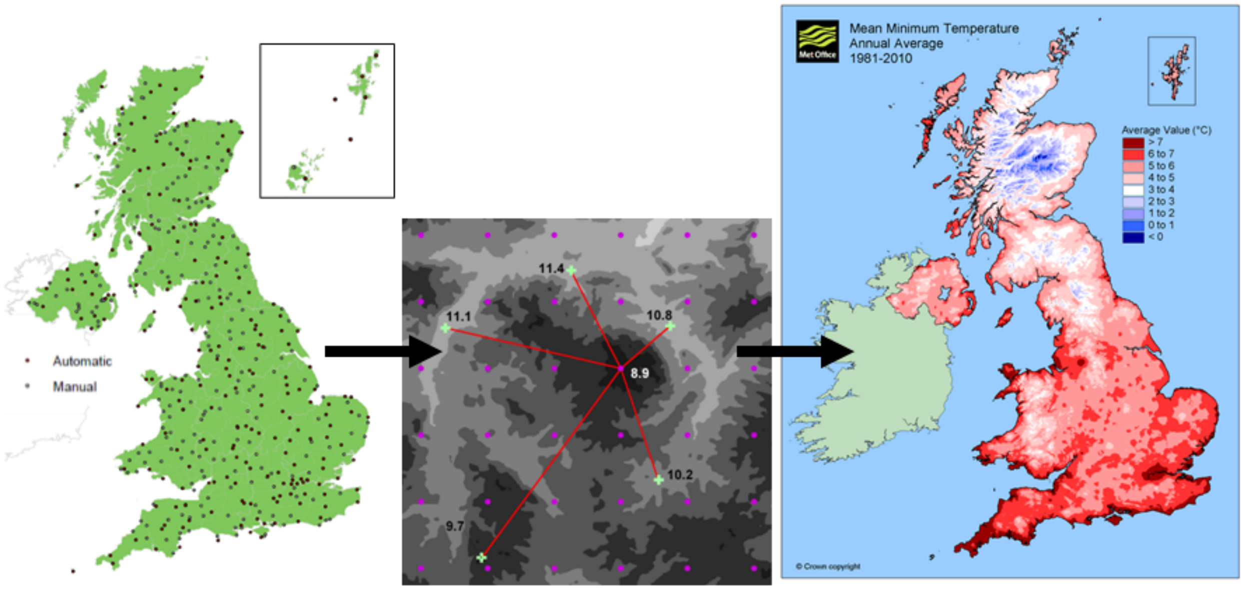
Some elements of the weather have strong diurnal and seasonal cycles, such as temperature that are intuitively understandable and can be readily accounted for within statistical analysis. For example for routine climate monitoring variability is usually presented as anomaly departures from a long term average. The standard long term average periods are thirty years and following World Meteorological Organisation guidance have been calculated for each calendar month for the periods 1961 1990, 1971 2000, and 1981 2010. For temperature the seasonal cycle is large enough that the coldest July on record for the UK is still higher than any month from October to May. For rainfall this is not the case, for example June is climatologically one of the driest months for the UK, but in 2012 was second only to December for the amount of rain that fell.

You can explore local and regional climate and seasonality here.

Back to flow chart

Back to Data Summary table

2.2.2.3 Sampling - representativity



The network of meteorological observations is traditionally designed to provide optimum representation of synoptic scale meteorology, such as the passage of weather systems, and therefore will generally (but not always) avoid localised micro-climates such as frost hollows, or dense urban environments. Urban microclimates are also highly variable, and some urban stations are located on roof tops, so caution is required for using and interpreting these data. If using station observations to infer weather or climate at a location some distance from a weather station the user should consider how representative the site might be. Are there other topographic features between the locations that might influence the results, for example locations geographically close but windward or leeward of mountains may have markedly different climates.

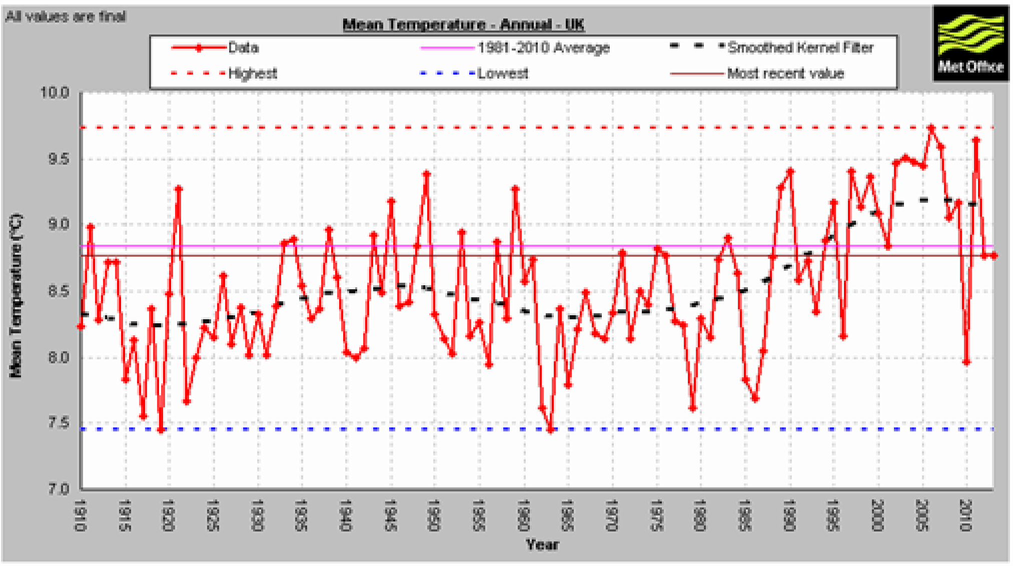
If your interests are in national or regional scale weather and climate the Met Office adopt a peer-reviewed method for handling the coverage and changes over time of the irregularly distributed observation network to derive robust estimates of spatially aggregated data.

To explore in more detail methods of spatial analysis visit our methods pages here.

Back to flow chart

Back to Data Summary table

2.2.2.4 Trends and autocorrelation



Meteorological data exhibit trends across a range of timescales arising from for example large-scale fluctuations such as the North Atlantic Oscillation that relates to pressure anomalies over the North Atlantic influencing weather patterns over the UK particularly during winter. There is also strong autocorrelation in some meteorological parameters out to inter-annual timescales, which should be considered in determining the significance of correlations with other types of data, particularly where those data also exhibit trends over time.

Back to flow chart

Back to Data Summary table

2.2.2.5 Dependent variables

Different meteorological parameters should not necessarily be considered independent within statistical modelling. The relationship between different climate parameters will also be seasonally dependent. For example warm summers will tend to be sunny and dry, in contrast a mild winter will tend to be stormy and wet. This example is explored in section 3.2.

It is also important to understand the mechanisms resulting in particular outcomes. For example in "Assessing the potential impact of climate change on the UKs electricity network" (link) while snow faults were identified as a significant contributor to weather-related network faults, there was little correlation with snow depth observations, this is because the faults are actually associated with a combination of snow and high wind gusts resulting in ice accretion on lines. Without an appropriate level of understanding of the underlying processes it would be easy to misinterpret or provide misleading advice.

Back to flow chart

Back to Data Summary table

2.2.3 Data types

The guide provides an overview of the different types of data, and potential purposes for which they can be used. It may be possible to develop other types of dataset or for guidance or enquiries contact Met Office for further information.

2.2.3.1 Climate summaries

Description:

Met Office produce daily, monthly, seasonal and annual weather and climate summaries. These are produced routinely within the first two to three working days of each calendar month. They provide a commentary overview of the previous months weather, key statistics, visualisations and place it into a historical context. In addition the Met Office publish more in depth discussion and analysis of extreme events, usually within a few weeks of their occurrence. These reports provide meteorological and historical context along with key observations associated with notable weather events affecting the UK.

Daily weather reports are available for every day since 1860 from the National Meteorological Library and Archive (link).

Monthly weather summaries are available for every month 1884-1993, and 2001- present.

Purpose:

The Met Office weather and climate summaries provide an authoritative contextual description of UK weather and climate, based on available observations at time of publication. They are a component of our obligation to provide a national memory of weather. The contents will be of particular value for those requiring commentary information and latest statistics.

How are they produced?

Written by Met Office staff based on Climate Series, Station Climate Statistics, and Daily Climate Observations.

Key benefits:

- Accessible formats
- Timely production
- Open access
- Multi-media

Access:

- Format: Text and images
- Delivery: Web

- Licence: Open Government Licence

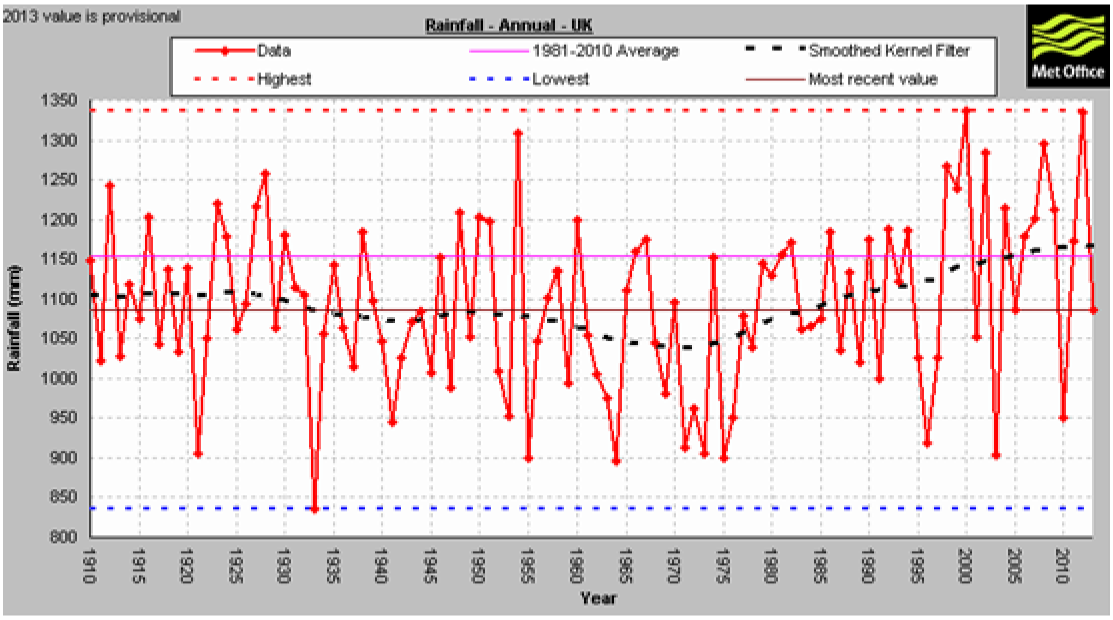
Example content:

- Daily weather reports
- Winter storms 2014
- Last month

Back to flow chart

Back to Data Summary table

2.2.3.2 Climate series



Description:

Met Office produce a number of long series homogeneous national and regional climate statistics derived from station observations. Further details of these data and their production are available here.

National and regional statistics
Monthly, seasonal and annual statistics derived from gridded climate datasets. These provide estimates of national and regional climate for a range of core climate variables from 1910 to present using all available observations. Updated monthly.

HadCET
Daily and monthly temperatures representative of a roughly triangular area of the United Kingdom enclosed by Lancashire, London and Bristol. The monthly series, which begins in 1659, is the longest available instrumental record of temperature in the world. The daily series begins in 1772.

HadUKP
UK regional precipitation, which incorporates the long-running England & Wales Precipitation (EWP) series beginning in 1766, the longest instrumental series of this kind in the world.

Purpose:

- Monitoring UK climate variability and change.
- Public climate records
- Academic Research

How are they produced?

Derived from Gridded climate data and daily climate observations methods described here.

Key benefits:

- More than 100 years of climate data
- Homogeneous series
- Simple and widely used
- Small data volume
- Open access
- Routine updates

Access:

- Format: Date or rank ordered statistics for months, seasons, and years.
- Delivery: Web
- Licence: Open Government Licence

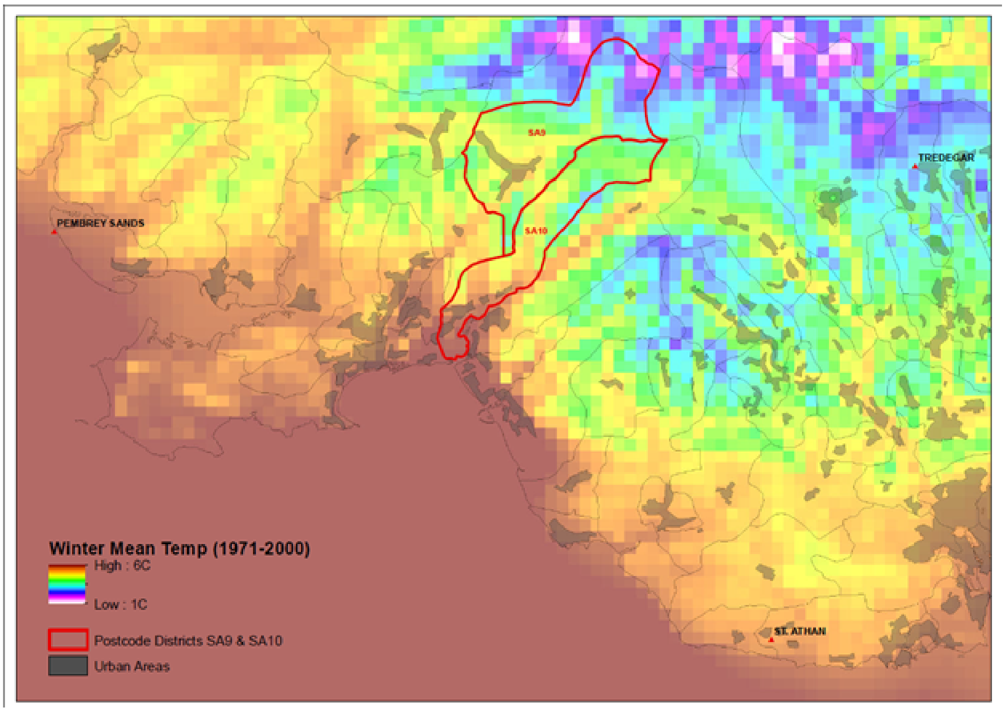
Example application:

- Retail sales analysis

Back to flow chart

Back to Data Summary table

2.2.3.3 Gridded climate data



Description:

The number, distribution and location of stations have changed substantially over time, and any individual station record can be prone to data gaps or errors. Furthermore users of climate information are often particularly interested in the ability to produce climate histories for areas such as administrative regions or postcode districts. Therefore there is a requirement to produce estimates of climate variables that are free from gaps in space and time. Climate variables depend on a variety of geographic and topographic factors, e.g. air temperature usually decreases with altitude but increases in urban areas. Consequently the estimated value of a variable at a location depends on more than just an interpolation of the values at the surrounding weather stations. The approach used by the Met Office is to:

- Adjust the observations in order to remove the effects of topography, using either anomalies from the long-term average and/or regression analysis.
- The resulting quantity is assumed to vary smoothly and can be interpolated using standard techniques.
- The effects of geography and topography at each grid point are then re-introduced, through either the long-term average or the regression model

The outcome is an interpolation of climate observations onto a uniform grid. Grid text files provide a matrix of values covering the whole of the UK, in one file for each day, month or year. Each value represents an estimate for the centre point of a 5 x 5 km grid cell. These grid cells are identified using the Ordnance Survey National Grid, extended to cover Northern Ireland. The area covered is shown in this map, with climate variable values for all land areas coloured green and -9999 elsewhere. For more information see here.

Purpose:

- Monitoring UK climate variability and change.
- Academic Research
- GIS analysis
- Can be used to generate weighted statistics e.g. temperature weighted by population or by crop distributions.

How are they produced?

Derived from Station climate statistics and Daily Climate Observations methods described in more detail here.

Key benefits:

- More than 100 years of data
- Data are complete in space and time
- Routine updates

Access:

- Format: Matrix of values covering whole of UK on a 5 x 5 km grid on the Ordnance Survey National Grid.
- Delivery: Web
- Licence: Non-commercial Government Licence

Example application:

- Retail sales analysis
- Gridded observations

Back to flow chart

Back to Data Summary table

2.2.3.4 Station climate statistics

Description:

Climate stations have historically made daily measurements at typically 0900 GMT, although there are some exceptions. These provide observations to meet the requirements of climate monitoring for the UK. Some of the network comprise Met Office automated stations, but a significant proportion is still provided unpaid enthusiasts making daily manual observations..
Some of the key observed elements from climate stations are listed below, but not all stations will have all the equipment to make all observation types, some may be limited to e.g. air temperature or rainfall.

- Maximum air temperature 0900GMT to 0900GMT the next day
- Minimum air temperature 0900GMT to 0900GMT the next day
- Air temperature at 0900GMT
- Grass minimum temperature dusk to 0900GMT the next day
- Soil temperature at 0.1m, 0.3m, 1.0m
- Relative humidity or wet bulb temperature equivalent at 0900GMT
- Rainfall amount 0900GMT to 0900GMT the next day
- Depth of snow lying
- State of ground
- Sunshine duration

The observing period is therefore an important consideration for users of the daily climate archive. A maximum temperature thermometer read at 0900GMT on day D is most likely to be providing the maximum temperature reached in the afternoon of the preceding calendar day. In contrast minimum temperatures are most likely to be reached just before dawn, therefore on the calendar day of the reading. Similarly most of the rainfall accumulation would have been in the preceding day. Therefore for most applications it is standard practice to throw back the Tmax and rainfall readings by one calendar day. The meteorological data archive however will store them at the end of the observing period to which they relate.

Purpose:

- Monitoring UK climate variability and change.

How are they produced?

As observed.

Key benefits:

- Direct observations from specific site locations.

Access:

- Terms and conditions apply. For academic and non-commercial government use the historical climate archive is available from the British Atmospheric Data Centre for other applications contact the Met OfficeMet Office.

Example application:

- Met Office Integrated Data Archive System (MIDAS)

Back to flow chart

Back to Data Summary table

2.2.3.5 Daily climate observations

Description:

Climate stations have historically made daily measurements at typically 0900 GMT, although there are some exceptions. These provide observations to meet the requirements of climate monitoring for the UK. Some of the network comprise Met Office automated stations, but a significant proportion is still provided unpaid enthusiasts making daily manual observations.

Some of the key observed elements from climate stations are listed below, but not all stations will have all the equipment to make all observation types, some may be limited to e.g. air temperature or rainfall.

- Maximum air temperature 0900GMT to 0900GMT the next day
- Minimum air temperature 0900GMT to 0900GMT the next day
- Air temperature at 0900GMT
- Grass minimum temperature dusk to 0900GMT the next day
- Soil temperature at 0.1m, 0.3m, 1.0m
- Relative humidity or wet bulb temperature equivalent at 0900GMT
- Rainfall amount 0900GMT to 0900GMT the next day
- Depth of snow lying
- State of ground
- Sunshine duration

The observing period is therefore an important consideration for users of the daily climate archive. A maximum temperature thermometer read at 0900GMT on day D is most likely to be providing the maximum temperature reached in the afternoon of the preceding calendar day. In contrast minimum temperatures are most likely to be reached just before dawn, therefore on the calendar day of the reading. Similarly most of the rainfall accumulation would have been in the preceding day. Therefore for most applications it is standard practice to throw back the Tmax and rainfall readings by one calendar day. The meteorological data archive however will store them at the end of the observing period to which they relate.

Purpose:

- Monitoring UK weather and climate.

How are they produced?

As observed.

Key benefits:

- Direct observations from specific site locations.

Access:

- Terms and conditions apply. For academic and non-commercial government use the historical climate archive is available from the British Atmospheric Data Centre for other applications contact the Met Office.

Example content:

- Met Office Integrated Data Archive System (MIDAS)

Back to flow chart

Back to Data Summary table

2.2.3.6 Hourly observations

Description:

Observations are made primarily for the purpose of providing information on the present state of the atmosphere for weather forecasting and are termed synoptic. A range of observations are made of meteorological variables at hourly resolution. Most synoptic observations are now automated, but some trained observers still operate particularly for some features that are harder to automate like visibility or cloud type that will be important for e.g. aviation.

Purpose:

- Monitoring state of atmosphere

How are they produced?

As observed.

Key benefits:

- Direct observations from specific site locations at sub-daily resolution.

Access:

- Terms and conditions apply. For academic and non-commercial government use the historical climate archive is available from the British Atmospheric Data Centre for other applications contact the Met Office.

- Real time observations are available under open government licence from the Met Office Datapoint service.

Example content:

- Met Office Integrated Data Archive System (MIDAS)

- Met Office Datapoint

- Trends and periodicities in extreme hourly rainfall CONVEX project

Back to flow chart

Back to Data Summary table

3 Initial data analysis

It is advisable to spend some time on your initial data analysis to get an idea of data quality and structure and to help inform any model building. Some of the initial data analysis may also be of interest to users in its own right.

3.1 Assessing data quality and structure

Some of the issues on data quality and structure will have been covered to some extent when selecting the data to be used in the analysis. Assessing the data quality and structure forms part of the initial data analysis and is a critical part of the process before starting to model the data. If it is not done and you miss some important aspects of the data you are using then this could lead to meaningless results and misinterpretation of your time series and any relationships to weather or climate data.

From your data selection you should have a good idea about the span of time covered and the frequency and coverage of the time series. It is worth spending some time to understand how the data have been collected, and whether there are particular events that may have impacted the time series. For example, changes in definitions, known anomalies or outliers, abrupt changes in the level or changes in seasonal pattern. Some of this sort of information will be obvious from plotting each time series, while other effects may need some further investigation. Some software provides algorithms for the automatic detection of certain types of outlier. Care should be taken when using these to ensure that what is identified makes sense. You may end up with a series that is modelled with lots of outliers and ends up with a poorly specified model and invalid inference about relationships. Knowledge of the series is crucial for identifying and appropriately including intervention variables.

In terms of understanding data structure from a time series perspective, one issue requiring special attention is the correlation structure of a time series. In a straightforward ordinary least squares regression model it is assumed that the observations are independent. In time series this assumption is often not valid and therefore needs to be addressed. Some useful time series tools for assessing this structure, such as the autocorrelation function are introduced in section 4.1.

Back to flow chart

3.2 Further preliminary analysis

Some initial data analysis has been suggested in the section 3.1. In this step the idea is to have some further exploration of how different time series might be related to one another. Simple time series plots of data can be a very powerful tool to help you think about potential relationships and will be important for specifying your model. In addition they can help to identify anomalies and structural changes in time series. Various summary statistics of your data may also be useful.

If the data have already been seasonally adjusted or if you already have a time series model that does not include weather effects, you could examine plots of the residuals or the irregular component from seasonal adjustment against weather type variables to get an indication of any possible relationships. For some weather variables, as the relationship may differ for different months, quarters or other periods it can be useful to examine plots of these relationships at different periods. For example, there may be a negative correlation between sunshine hours and temperature in the winter months and positive correlation in summer months. Figure 2 shows a scatter plot of deviations from monthly average maximum daily temperature against deviations from monthly average sunshine hours. These are UK data covering the span January 1986 to April 2014. A line of best fit suggests some positive correlation but the correlation does not appear to be very strong (0.45).

Figure 2: *Scatter plot of deviations from monthly average maximum daily temperature against deviations from monthly average sunshine hours.*

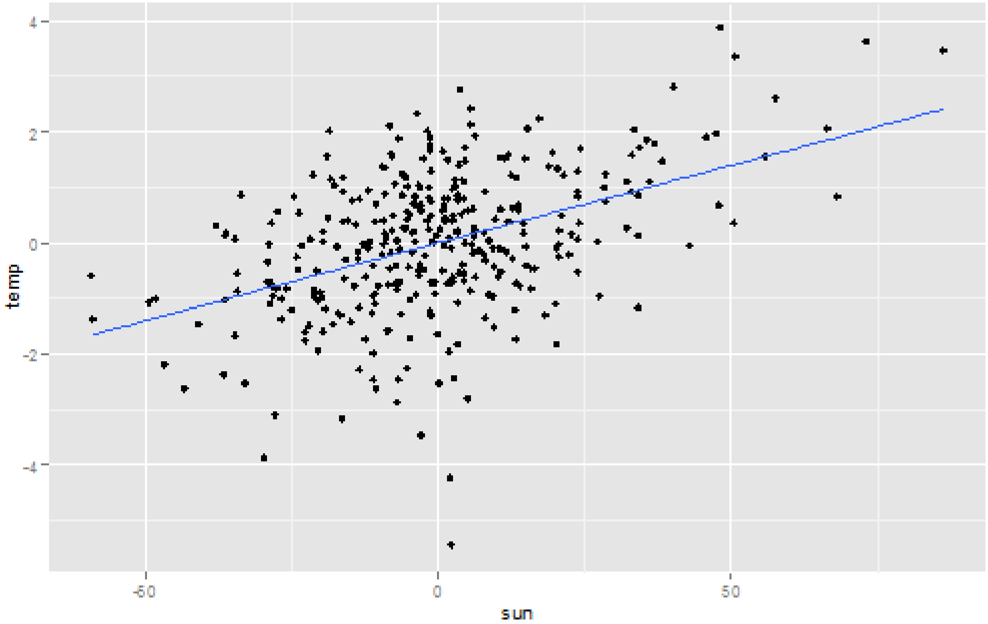
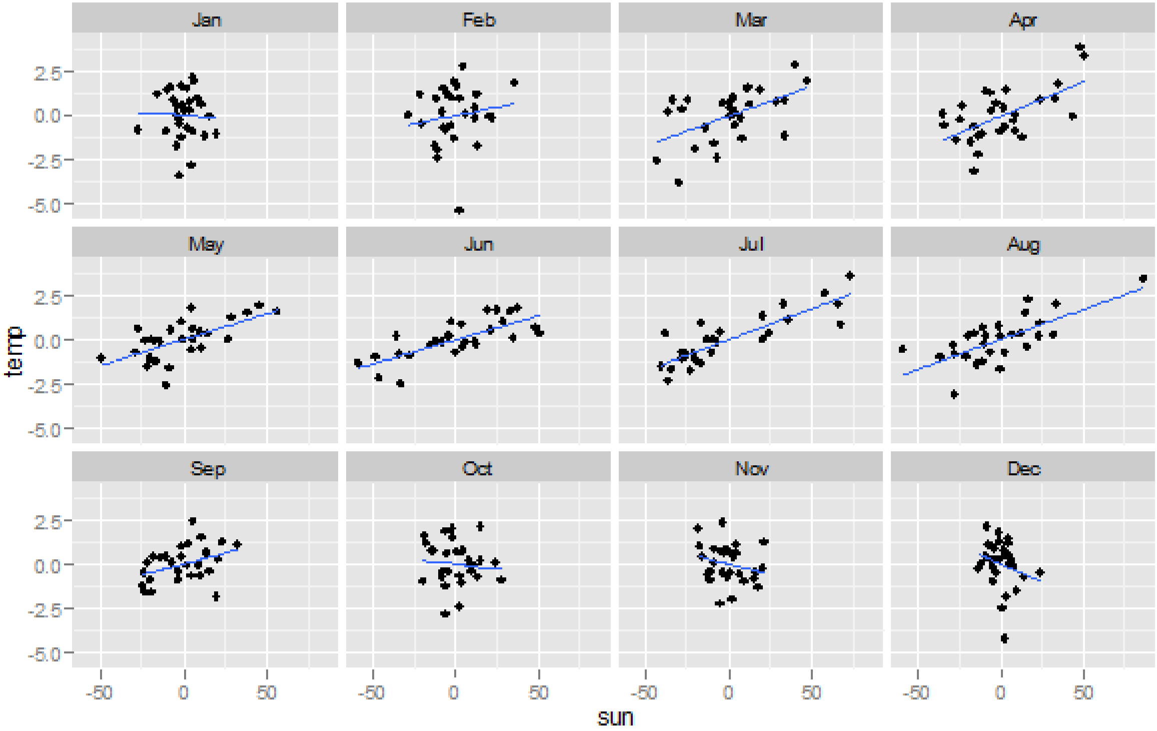


Figure 3 shows exactly the same data but arranged by month. The positive correlation is more evident in the summer months (for example, 0.85 for July) and less apparent in the winter months (for example, -0.24 for December). Given the evidence of different relationships for different months this gives us useful information for constructing our model. We are likely get an improved model by modelling the relationship between temperature and sunshine hours by month. Figure 3.2: Scatter plot of deviations from monthly average maximum daily temperature against deviations from monthly average sunshine hours by month.

Figure 3: *Scatter plot of deviations from monthly average maximum daily temperature against deviations from monthly average sunshine hours by month.*



Back to flow chart

4 Time series modelling

This section provides a very brief introduction to time series analysis in order to provide readers with some familiarity of concepts discussed in the examples. It is not intended to be in any way a comprehensive introduction to time series analysis and users of this guide are strongly encouraged to follow the references for further information.

4.1 Time series concepts

What is a time series? A time series is a sequence of observations of the same phenomenon over time. In official statistics we often deal with more or less regularly spaced observations in time, such as monthly retail sales, or annual crop yields. These are examples of discrete time series as opposed to a continuous measurement of a variable over time; our examples are all discrete time series.

Flow or stock? Time series are sometimes further categorised as either flow or stock data; for example, total turnover in a month calculated as the sum of daily turnover over the month is a flow variable, whereas the number of people employed on the last day of the month is a stock variable. The examples of time series modelling included in this guide deal with flow data, but a similar approach to modelling can be used for stock data, although care needs to be taken in building an appropriate model, for example if your stock is measured on a particular day each month, it may not be appropriate to use a weather or climate variable measured as some average over the month.

What is time series analysis? Time series analysis can involve explaining movements in a time series for different purposes such as understanding current or historical movements, or interpreting relationships with other time series. These explanations may be used for inference and prediction. The range of time series methods for analysis is wide and this guide is very specific focusing on ARIMA models with exogenous regression variables (mostly weather and climate data in the examples). Accessible introductions to time series analysis are given in (Chatfield, The Analysis of Time Series: An Introduction Sixth Edition, 2004), (Shumway & Stoffer, 2006) and (Harvey, 1993).

Seasonal adjustment A widely used method of time series analysis in official statistics is seasonal adjustment. This involves decomposing an observed time series into a trend, seasonal and irregular component. The main purpose of this is to remove the seasonal component to aid interpretation of movements in monthly or quarterly time series. Further information on seasonal adjustment can be found in (ONS, 2014), (USCB, 2013), (Ladiray & Quenneville, 2001), and (Gomez & Maravall, 2001).

The headline figures published in monthly or quarterly releases are usually changes in the seasonally adjusted series. Statisticians responsible for these publications provide commentary on the seasonally adjusted time series and users are often interested to understand why there are particular movements in series.

Part of the process of seasonal adjustment usually involves modelling a time series to deal with some often observed properties of a time series. In 2012 the GSS Statistical Policy and Standards Committee recommended using a program called X-13ARIMA-SEATS for the seasonal adjustment of UK official statistics. This is freely available software produced at the United States Census Bureau (USCB, 2013). While the main purpose of X-13ARIMA-SEATS is seasonal adjustment, it also includes time series modelling capabilities in the form of what are termed regARIMA models (ARIMA models with exogenous regression variables).

Back to flow chart

4.2 Time series models

There are many different time series models that could be explored. In this guide we focus on the regARIMA model that is used in the GSS recommended software for seasonal adjustment, X-13ARIMA-SEATS.

ARIMA models (autoregressive integrated moving averages) are well established time series models that can deal with the autocorrelation structure in a time series. The term regARIMA is used to describe the combination of a regression model and an ARIMA model.

This section provides a simple overview of the regARIMA model and describes some simple steps for specifying and estimating the model using X-13ARIMA-SEATS. The practical examples in this guide include data and code for building such models in X-13ARIMA-SEATS, R and SAS. Some further information on a practical guide to regARIMA modelling is provided in the ONS Guide to Seasonal Adjustment (ONS, 2014). A thorough description of the regARIMA modelling capabilities of X-13ARIMA-SEATS is provided in chapters 4 and 5 of the softwares reference manual (USCB, 2013). For an introduction to ARIMA models see (Chatfield, 2004).

Users of this guide are strongly encouraged to consult a standard time series analysis text book to get an understanding of basic time series terminology and issues. Here we provide a very brief introduction to some important time series concepts that will be discussed in the modelling examples.

Back to flow chart

4.2.1 The regARIMA model

The regARIMA model can be thought of as a linear regression model with ARIMA errors. It takes the form of equation 10.

$$y_t = \sum_{i=1}^{m} \beta_i x_{it} + z_t \tag{1}$$

where y_t is the time series of observations you want to model with observations at time points $t = 1, \ldots, T$. x_{it} are the $i = 1, \ldots, m$ regression variables that you want to include in your model. These might include outlier variables, calendar effects, and importantly for this guide weather or climate variables. β_i are the regression coefficients to be estimated in the model. z_t is an ARIMA process, which for now we simply note deals with certain time series characteristics that could lead to poor inference in our regression model if not appropriately dealt with.

For example, assume we have a monthly time series y_t and a variable x_{it} that measures the difference between the observed average maximum daily temperature in month t and the long term average maximum daily temperature for that month (for example the monthly averages averaged over 28 years). The estimated coefficient β_i can be interpreted as the amount you would expect y_t to change if the average maximum daily temperature in any month is one degree above the long term average, holding everything else constant. This model assumes that the relationship between temperature and the time series of interest is constant across months, which may not be the case, and the examples in this guide show variables that deal with this issue.

It is also worth noting that the time series y_t may be transformed by some function. A common transformation for many official statistics time series is to take a log transformation. This alters the interpretation of the estimated coefficients. Using the previous example and assuming that y_t is a log transformation of the variable of interest the estimated coefficient is an estimate of the per cent change of the series of interest for an average maximum daily temperature that is one degree above the long term average.

In terms of some commonly used terminology, y_t is described as a stochastic process. This stochastic process in the model above is the sum of a deterministic part (the regressors x_{it}, which we assume are not random) and a stochastic part z_t that deals with the problem that errors in a time series may be correlated. y_t is described as a stochastic process as it is a collection of random variables at different points in time, where the random element comes from the stochastic part z_t which is a collection of realizations from some random distribution.

In summary, the regARIMA model includes the following.

- A transformation of the series of interest (may be no transformation).

- A regression model where we will include the weather and climate variables.
- An ARIMA model that deals with some time series characteristics that may cause problems for inference in a standard regression model.

Back to flow chart

4.2.2 Stationarity

If we have a stochastic process y_t, this is described as stationary if the mean, and covariance properties of the series are not time dependent. For example, a series that has an increasing trend would be non-stationary, as the mean or average level of the series is changing with time. The four plots in figure 4 show some examples of stationary and non-stationary time series. Series 1 has an upward trend and so the mean is time dependent although the variance is not time dependent. Series 2 has a mean and variance that are time dependent and so is also non-stationary. Series 3 and 4 both have a mean and variance that are not time dependent and are both stationary, although they have a different correlation structure as discussed below.

In an ARIMA model, the "I" part (integrated) and sometimes a transformation are used to make the series stationary before dealing with the correlation in the AR (autoregressive) and MA (moving average) parts (these are explained in more detail in section 4.2.6). For example, the linear upward trend in series 1 would be removed by taking the first difference of the time series $y_t - y_{t-1}$, whereas the variance of series 2 may be stabilised with a log transformation. Note that it also requires differencing due to the upward trend. A straightforward time plot of series 3 and series 4 in figure 4 does not reveal much about the structure of either series. They could both appear to be noise around a constant mean of 100. Both of these series are stationary. However, series 3 was simulated from an autoregressive process, whereas series 4 is 120 random draws from a normal distribution.

Back to flow chart

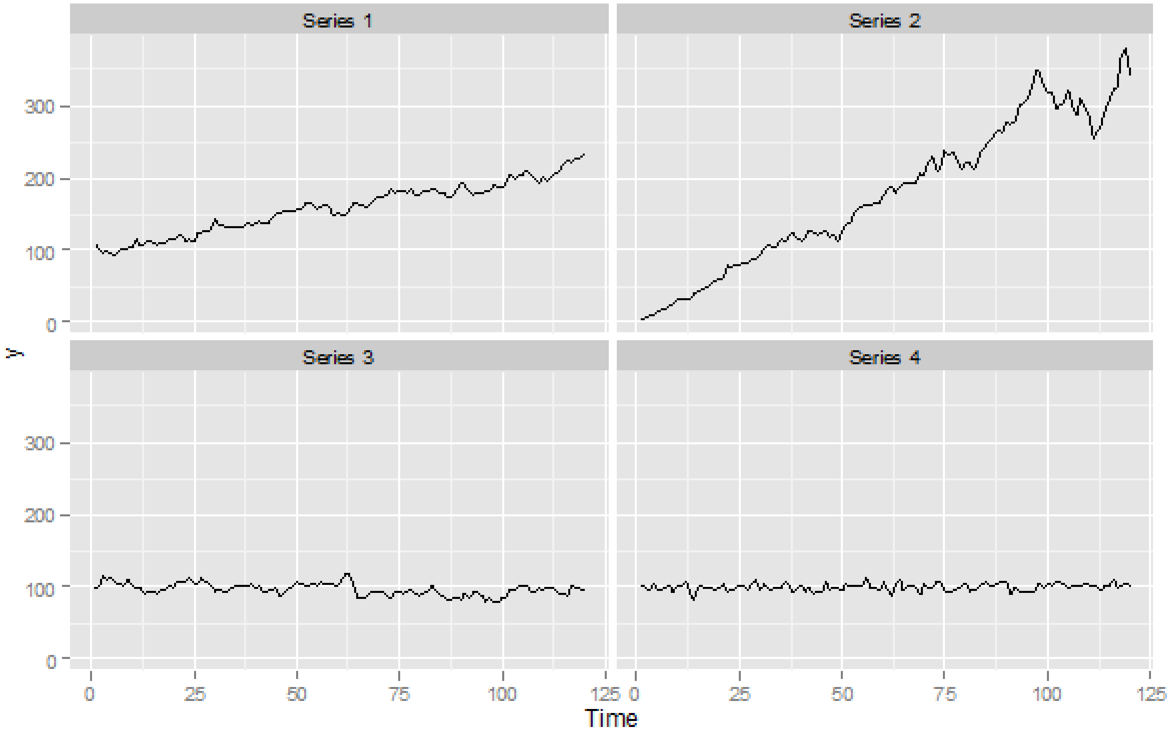
4.2.3 Autocorrelation

A useful plot for time series analysis is the autocorrelation function (ACF) which shows the correlation at different lags. The ACF of the series plotted in figure 4 are shown in figure 5.

Series 1 and 2 have not been differenced and therefore the trend (non-stationarity) will cause the autocorrelation function to die away slowly and still show positive correlation at high lags. When attempting to specify an ARIMA model and work out the order of differencing an ACF such as those for series 1 and 2 suggests that differencing is required.

The ACF for series 3 shows a large positive correlation a lag 1. Note that the correlation at lag 0 is 1 by definition, but that the correlation is reducing for greater lags. Such autocorrelation, if found in your model residuals, needs to be dealt with, or it will lead to poor inference. The ACF for series 4 shows no autocorrelation in the series. If an

Figure 4: *Examples of stationary and non-stationary time series (top row series are non-stationary and bottom row are stationary)*

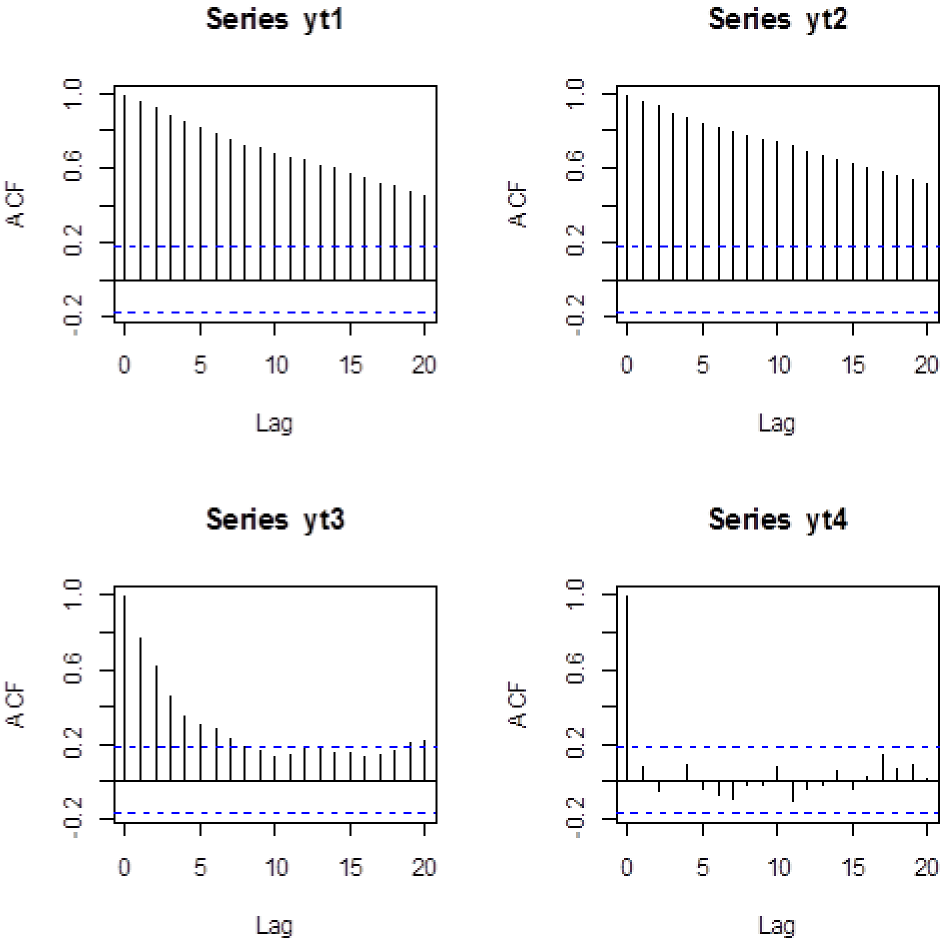


ACF plot of your model residuals looked like that for series 4, there is no evidence of autocorrelation in the residuals.

Depending on the nature of the correlation structure, this can be effectively dealt with by the ARMA part of the model (remember differencing (the “I” part of ARIMA) should already have been done if required to make the series stationary). The ACF and the associated partial autocorrelation function can be used to help identify the orders of an ARIMA model. Further information on the use of these for manual identification of an ARIMA can be found in section 4.1.1 of (Chatfield, 2004).

Back to flow chart

Figure 5: *Estimates of the autocorrelation function (ACF) for series 1 to 4*

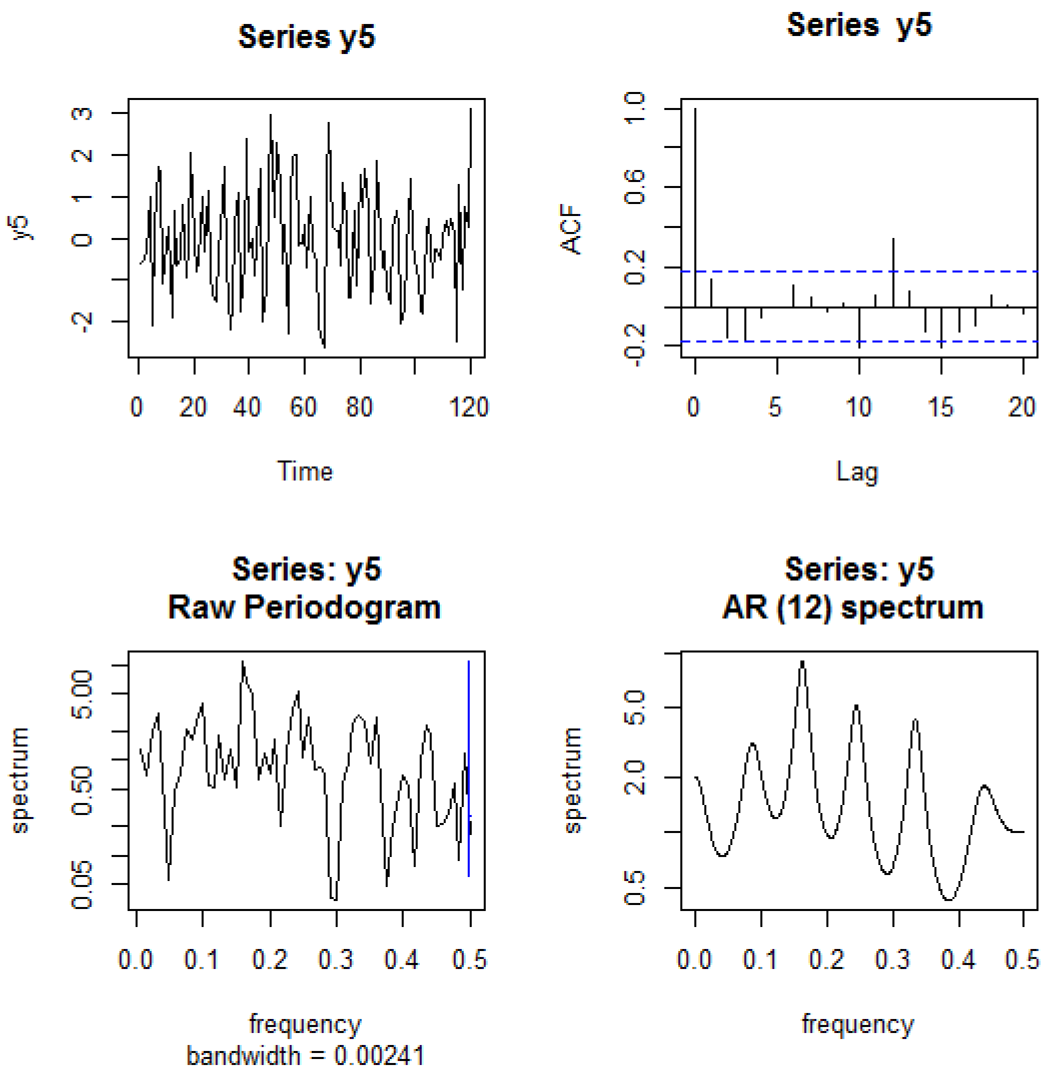


4.2.4 Spectral plots

Another useful and related tool for analysing the structure of a time series is the spectrum, shown in spectral plots. This contains the same information as the ACF but is presented in a different way and provides a useful complementary tool. While some of the maths may be quite advanced, there is still practical use in interpreting estimates of the spectrum. It is used as one of the diagnostic tools in X-13ARIMA-SEATS to help identify residual seasonality. It shows the contribution to the variance of a time series at different frequencies. These spectral plots show some measure of the contribution to the variance of the series on the y-axis and frequencies on the x-axis. Sometimes the labels can differ as they show slightly different representations of the same information.

The spectral plot can be used to identify patterns in the data. This can be useful when assessing your model, as your residuals should not show any pattern. For example, figure 6 shows different types of plots for a time series of simulated model residuals (series y5); the upper left is a straightforward time series plot, upper right is the ACF, the lower left shows the raw periodogram, and the lower right shows an alternative estimate of the spectrum.

Figure 6: *Time series plot (upper left) , ACF (upper right), periodogram (lower left) and spectrum (lower right) of simulated model residuals (series y5)*

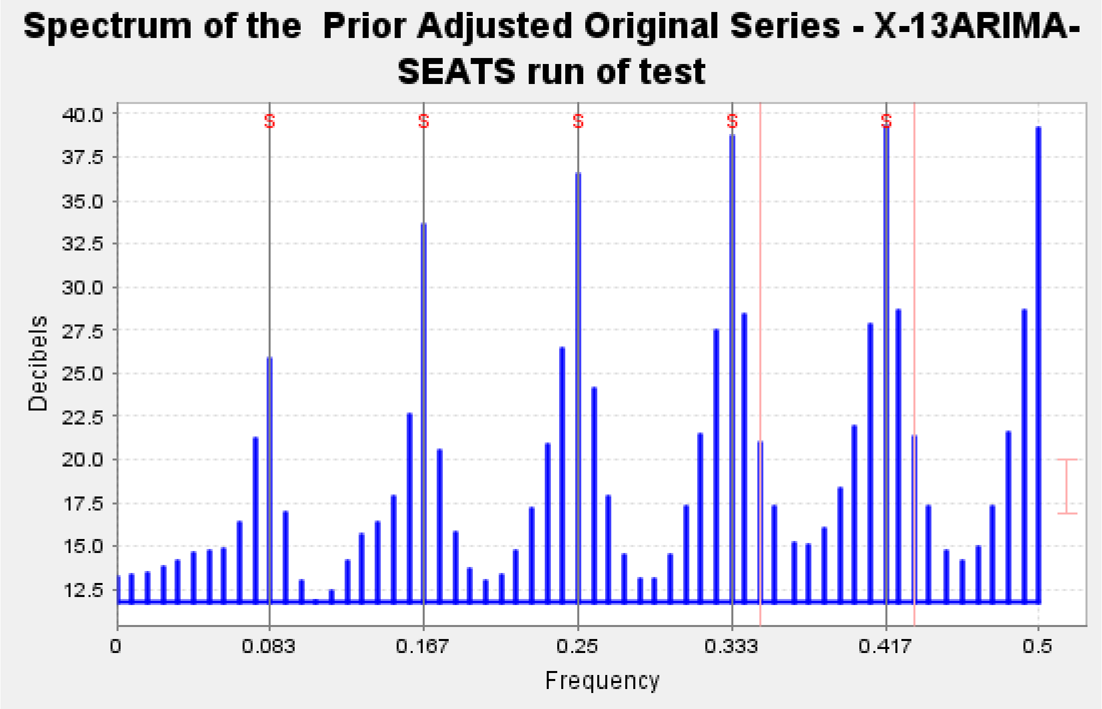


The spectral plot shown in the bottom right is created by modelling the series as an autoregressive process of order 12 and computing the spectral density function of the fitted model. The effect of this is a smoother estimate than shown in the periodogram. For more information on the range of methods used for estimating the spectrum see chapters 6 and 7 of (Chatfield, 2004). The spectrum calculated by default in X-13ARIMA-SEATS uses the estimated coefficients from an autoregressive model of order 30. For more information on the spectral plots produced by the software see chapter 6.1 and chapter 7.17 of (USCB, 2013).

The spectrum in figure 6 shows evidence of seasonality; this is evident from the peaks that appear at what are known as seasonal frequencies. This evidence of residual seasonality is also shown in the ACF with significant correlation at lag 12 (the peak greater than the 95% confidence interval that is based on a standard error calculated as 1/T where T is the length of the series). The spectrum plotted in X-13ARIMA-SEATS usefully includes vertical lines to show you seasonal frequencies and also trading day frequencies (cyclical patterns that may occur in data caused by the arrangement of the calendar) to help the user identify problems.

An example of a spectral plot from X-13ARIMA-SEATS is shown in figure 7. The vertical lines with “S” at the top are seasonal frequencies. There are also two other vertical lines at 0.348 and 0.432. These are trading day frequencies for monthly series and peaks at these frequencies in model residuals may indicate it is appropriate to include a trading day regressor or regressors in your model. Peaks at seasonal frequencies may also indicate that a change to the model is required to deal with seasonality better.

Figure 7: *Spectral plot produced by X-13ARIMA-SEATS show strong evidence of seasonality (peaks at seasonal frequencies)*



Back to flow chart

4.2.5 Further analysis of model residuals

There are other problems that may be evident from an analysis of model residuals, and appropriate analysis of model residuals is an important part of assessing your model. One other scenario that we do not really address in this guide is where you have changes in variance. We have mentioned the use of a log transformation to stabilise the variance and other transformations (Box-Cox transformations) may also be appropriate. As already noted, there are many different types of models in the literature of time series modelling, and some that specifically deal with changes in variance that could be explored if necessary, for example ARCH (autoregressive conditional heteroskedasticity) and other related models. State space models also provide a very flexible form of modelling but are not dealt with in this guide. For an introduction to some of these other models see for example (Chatfield, 2004) or (Harvey, 1993).

Back to flow chart

4.2.6 Examples of ARIMA models

What does an ARIMA model look like?

This section provides a short description of the components of an ARIMA model and introduces some notation used in the examples. The I part stands for integrated and we say a model is integrated of order d. For example, if $d = 1$, then a first difference is taken of the time series, and it is the first difference that is modelled as an ARMA process.

First differences would remove a linear trend, whereas as second differences would remove a quadratic trend. Equation 2 shows an example of taking second differences of a time series y_t.

$$z_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \tag{2}$$

If we were fitting an ARIMA model with no AR or MA components and only an “I” component of order $d = 1$ we would be fitting the model in equation 3.

$$y_t - y_{t-1} = \epsilon_t \tag{3}$$

Note that ϵ_t is assumed to be independent and identically distributed. Such models are usually written as an I(d) process, so I(1) for the equation above. In the literature on ARIMA models, differencing is often presented by means of a backshift or lag operator (B or L). We use the backshift operator ($By_t = y_{t-1}$) and therefore the I(d) model could also be written as in equation 4.

$$(1 - B)^d y_t = \epsilon_t \tag{4}$$

The AR part of the model we describe as being of order p. The autoregressive part describes how the current value is related to past values of the time series. For example, an AR model of order 2, that we write as an AR(2) process is written as in equation 5.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \tag{5}$$

And a model of order p is written as in equation 6.

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t \tag{6}$$

The MA part of the model we describe as being of order q. The moving average part of the model shows how the time series is related to past errors, or innovations. For example an MA model of order 2, that we write as an MA(2) process is detailed in equation 7.

$$y_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t \tag{7}$$

And a model of order q is written as in equation 8.

$$y_t = \theta_1 y_{t-1} + \ldots + \theta_q y_{t-q} + \epsilon_t \tag{8}$$

We have very briefly introduced the notation used for parts of an ARIMA model. A common notation for the combination of these models is an ARIMA(p, d, q) model. Sometimes this is extended to describe seasonal ARIMA models, sometimes described as SARIMA. A common notation for such models is $(p, d, q)(P, D, Q)_s$. In this notation the P, D, and Q are seasonal AR, I and MA parts, while s is the frequency (sometimes the s subscript is dropped). Therefore an $(0, 0, 0)(1, 0, 0)_s$ model, where $s = 12$, is simply as in equation 9.

$$y_t = \Phi_1 2 y_{t-12} + \epsilon_t \tag{9}$$

Often ARIMA models involve some transformation of the variable of interest. A common transformation is the log, although other transformations may be appropriate. Where a transformation is applied, y_t in the above equations would be the transformed series.

What does a regARIMA model look like?

A regARIMA model can be described as an ARIMA model with a time varying mean; the time varying part caused by the regressors. Or it can be thought of as a regression in which certain time series properties are appropriately dealt with in an ARIMA model of the errors. Therefore a general seasonal regARIMA model can be written as in equation 10.

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D(y_t - \sum_{i=1}^{m} \beta_i x_{it}) = \theta(B)\Theta(B^s)\epsilon_t \tag{10}$$

For those who are new to this notation the above equation may appear complex. In the equation we have introduced the backshift operator for the non-seasonal and seasonal AR and MA polynomials. These are defined as in equation 11.

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \ldots - \phi_p B^p \quad \text{this is the non-seasonal AR polynomial} \\ \Phi(B^s) &= 1 - \Phi_1 B^s - \ldots - \Phi_P B^P s \quad \text{this is the seasonal AR polynomial} \\ \theta(B) &= 1 - \theta_1 B - \ldots - \theta_q B^q \quad \text{this is the non-seasonal MA polynomial} \\ \Theta(B^s) &= 1 - \Theta_1 B^s - \ldots - \Theta_Q B^Q s \quad \text{this is the seasonal MA polynomial} \end{aligned} \tag{11}$$

If we wish to fit a regARIMA model with one regressor (m=1) then clearly the reg part is simply $y_t - \beta_1 x_{1t} = z_t$. If $p = 0, d = 1, q = 1, P = 0, D = 1$, and $Q = 1$ are set for the order of the ARIMA model assuming a monthly time series ($s = 12$), then in short notation we can describe this as an $(0, 1, 1)(0, 1, 1)_{12}$ and the full regARIMA model would be as in equation 12.

$$\begin{gathered} z_t = z_{t-1} + z_{t-12} - z_{t-13} - \theta_1\epsilon_{t-1} - \Theta_1\epsilon_{t-12} + \theta_1\Theta_1\epsilon_{t-13} + \epsilon_t \\ \text{As} \quad \phi(B) = 1, \quad \Phi(B^s) = 1, \quad (1-B)^d = (1-B), \quad (1-B^s)^D = (1-B^{12}), \\ \theta(B) = 1 - \theta_1 B, \quad \Theta(B^s) = 1 - \Theta_1 B^{12} \end{gathered} \tag{12}$$

In this model we are saying that the observation at y_t depends on the value of the regressor at time t and lags 1,12 and 13. It also depends on lagged values of itself again at 1,12 and 13, and lagged and current period innovations. However, in terms of interpreting the effect of a weather variable at time t on the variable of interest at time t, we are interested in the value of the estimated regression coefficient as we would be in a standard regression.

We have presented a brief outline of some important time series concepts that will be discussed in some of the examples to give users of this guide some familiarity and also to point them to further information. These concepts are important to help users think about how to construct an appropriate model of their series, including weather and climate data. This guide provides examples of how to specify, fit and evaluate a model using X-13ARIMA-SEATS, R or SAS®. Next we describe the general steps involved in specifying, fitting and evaluating a time series model that includes weather or climate variables as regressors.

How do we specify, fit and evaluate a model?

When building a time series model in any software you will probably follow four steps.

1. Format your data
2. Specify the model
3. Estimate the model
4. Evaluate the model

A brief description of these steps is provided here and code and data are provided in the examples to help you get started.

1. Format your data

Depending on the software that you are using you will need to either load the data into the software, for example if you are using R or SAS® or if you are using X-13ARIMA-SEATS then you will need to save the dependent variable (the one that you wish to model) in one file and any regressors that are not provided by the programme such as your weather or climate variables will need to be saved in another file. A brief description of setting up and running a regARIMA model in X-13ARIMA-SEATS, R and SAS® is provided in section 4.3.

2. Specify the model

Depending upon the software that you are using there will be different ways in which the model is specified. For the modelling we consider in this guide you will need to choose an appropriate transformation of the dependent data. This can be done manually, or in some software it can be done automatically. Some care should be taken when using automatic procedures for creating a model as they may not be able to deal with certain specific structures in your data. However, they can form a useful starting point. You will need to decide what regressors to include in your model. Again some software includes automatic procedures for including certain types of regressors; see for example chapters 4.3 and 7.13 of (USCB, 2013) and chapters 9 to 12 of (ONS, 2014). For the models considered in this guide the weather or climate data will be regressors. Finally the choice of ARIMA model required for dealing with the issues discussed above needs to be specified. The ARIMA parameters may be specified manually and some software also provides automatic procedures for model identification.

3. Estimate the model

Depending on the software used, there may be different options available for how the model is actually estimated. These issues are beyond the scope of an introductory guide, and users are unlikely to need to change defaults used in different software, but they should be aware that these options exist, especially where results for what appear to be the same model differ slightly in different software.

4. Evaluate the model

Different software will produce different diagnostics and test statistics that are useful for evaluating your model. As discussed above analysis of model residuals are important. These are discussed further in the practical examples. The model diagnostics will help you to assess any possible improvements or required changes to your model. If you have sufficient data, it is useful to fit your model to different spans of your data to see whether you get stable estimates of the relationships to weather or climate data that you are interested in. In the models considered in this guide we generally assume that the relationship (estimated by the regression coefficient) is constant over time. You may end up with a number of models that all seem reasonable and so comparison of different models will also be important. These issues are discussed further in the examples.

This process of specifying, fitting and evaluating a model will probably be iterative and can take considerable time.

Back to flow chart

4.3 Software

Code and data are provided for you to replicate the example analyses contained in this guide. The main focus of the guide is regARIMA modelling as discussed in section 4 and so the examples tend to focus on X-13ARIMA-SEATS, which is a program for seasonal adjustment but includes some time series modelling capabilities. For more general statistical software such as R or SAS® there are a number of ways in which these models may be fitted and we do not attempt to cover them all. The code presented does not constitute a recommendation for particular functions in R or SAS$^{®}$ and is presented only to help users of this guide get started with their own analysis. This chapter provides a very brief introduction to each of the software and provides some example code for getting started.

In the examples below we will use retail sales data on clothing as our variable of interest, which we will describe as y and the deviations from average monthly rainfall in which we will describe as x.

Back to flow chart

4.3.1 X-13ARIMA-SEATS

This software is freely available from the US Census Bureau. There are a number of ways in which to run X-13ARIMA-SEATS. In this guide we assume that you are using the user interface called Win X-13 developed by the US Census Bureau. However, the descriptions of the data and spec files given below are applicable for running the software in other ways as described in reference manual (USCB, 2013). The version of X-13ARIMA-SEATS used in these examples is version 1.1 build 9 unless otherwise stated.

X-13ARIMA-SEATS needs a spec file as a minimum input from the user. However, in the examples in this guide we will also save the data we require in data files. A spec file tells the program what to do with the data you provide in the data files. Below are the specifications to fit a simple model of x on a log transformation of y with a seasonal ARIMA model specified as an $(0,1,1)(0,1,1)_{12}$, see section 4.2.6 for a brief explanation of this notation.

The y data are saved in a file called "y_data.dat" (right click here to save file) that is saved in an "x12save" format. This is a tab delimited text file, where the first column has dates in the format shown below and the data in the second column. These data always start in the third row. Note this is only an extract of the data. The full data file is also available at the end of this section.

```
Date    y_data
------  -----------------------
198601  3.12E+01
198602  2.65E+01
198603  2.92E+01
198604  3.16E+01
...     ...
201304  9.65E+01
201305  1.03E+02
201306  1.08E+02
```

The x data are saved in a file called "x_data.rmx" (right click here to save file). This is saved in the same "x12save" format. Again it is a tab delimited text file, but the extension used is ".rmx".

The spec file, saved here as y_data.spc (right click here to save file) contains a number of specifications; a specification or spec is text followed by a curly bracket. The text between the curly brackets is the parameters for that particular specification. The series spec must come first, but the order of the remaining specs is not important. The whole spec file is given below. Note that any text on the same line after a # is commented out and will not be read.

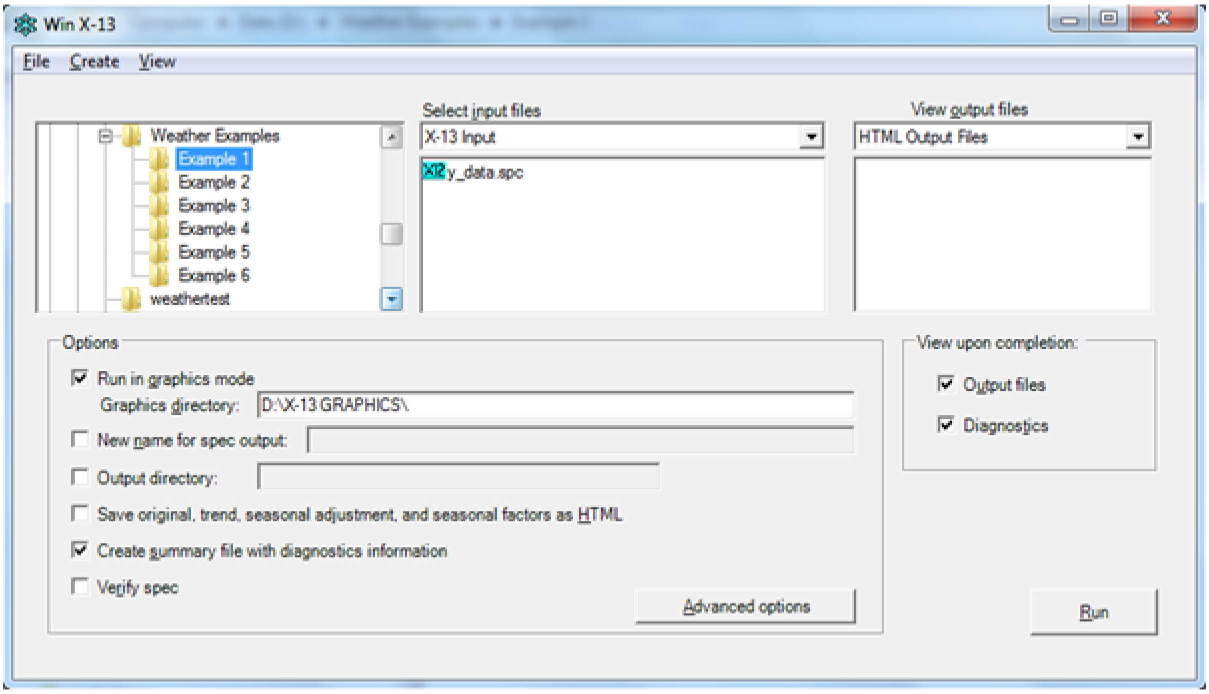
```
series{                #Start of the series spec
 file =``y_data.dat''  #Parameter of the series spec; file name for y data
 format =``x12save''   # Parameter of the series spec; format of y data
 period = 12           #Parameter of the series spec; period = monthly data
 }                     #End of the series spec
transform{             #Start of the transform spec
 function = log        #Parameter of the transform spec; log transform y
 }                     #End of the transform spec
arima{                 #Start of the arima spec
 model=(0 1 1)(0 1 1)  #Parameter of the arima spec; order of the ARIMA model
 }                     #End of the arima spec
regression{            #Start of the regression spec
 user=(x)              #Parameter of the regression spec; name for regressor x
 file=``x_data.rmx''   #Parameter of the regression spec; file name for x data
 format=``x12save''    #Parameter of the regression spec; format of x data
 usertype=user         #Parameter of the regression spec;
}                      #End of the regression spec
x11{}                  #Start and end of the x11 spec
```

There are a few points that are worth noting about the spec file. The file names for the y data and x data do not contain a full directory path; this will work only if the data are saved in the same directory as the spec file. There are alternative formats that can be used for the data. Note that the x11 spec is one of the available specifications for seasonally adjusting the data. For further information on setting up spec files in X-13ARIMA-SEATS and a full list of all specifications and parameters see chapters 3 and 7 of (USCB, 2013) and (ONS, 2014) for some practical examples.

With the data files and spec file specified, it is now possible to run X-13ARIMA-SEATS and generate output. If you are using Win X-13, in the upper left panel navigate to the directory where you have saved your files. In the upper middle panel you should get a list of all the spec files in that directory, as shown in the screenshot in figure 8. To run the spec file you can either double click on the spec file or use the Run button in the bottom right.

Note that in the screenshot above the option "Run in graphics mode" is selected and therefore as well as the standard output a number of graphs are also produced. Win X-13 produces a number of diagnostics saved in a table and each time you run or re-run a spec file the results are added to the diagnostic window. The output and appropriate charts should be examined to assess the model. The diagnostic window is useful especially for comparing different models. If you wish to save model diagnostics in the output then you will need to specify the appropriate parameters in the appropriate specifications. Further information on available diagnostics can be found in chapters 4.6 and 7 of (USCB, 2013). The examples in this guide also discuss a number of these diagnostics. The files used in the above example are given below.

Figure 8: *Screenshot of Win X-13*



Select input files: X-13 Input

y_data.spc

View output files: HTML Output Files

Options

☑ Run in graphics mode

Graphics directory: D:\X-13 GRAPHICS\

□ New name for spec output:

□ Output directory:

□ Save original, trend, seasonal adjustment, and seasonal factors as HTML

☑ Create summary file with diagnostics information

□ Verify spec

View upon completion:

☑ Output files

☑ Diagnostics

Right click here to save “y_data.dat” file
Right click here to save “x_data.rmx” file
Right click here to save “y_data.spc” file

Back to flow chart

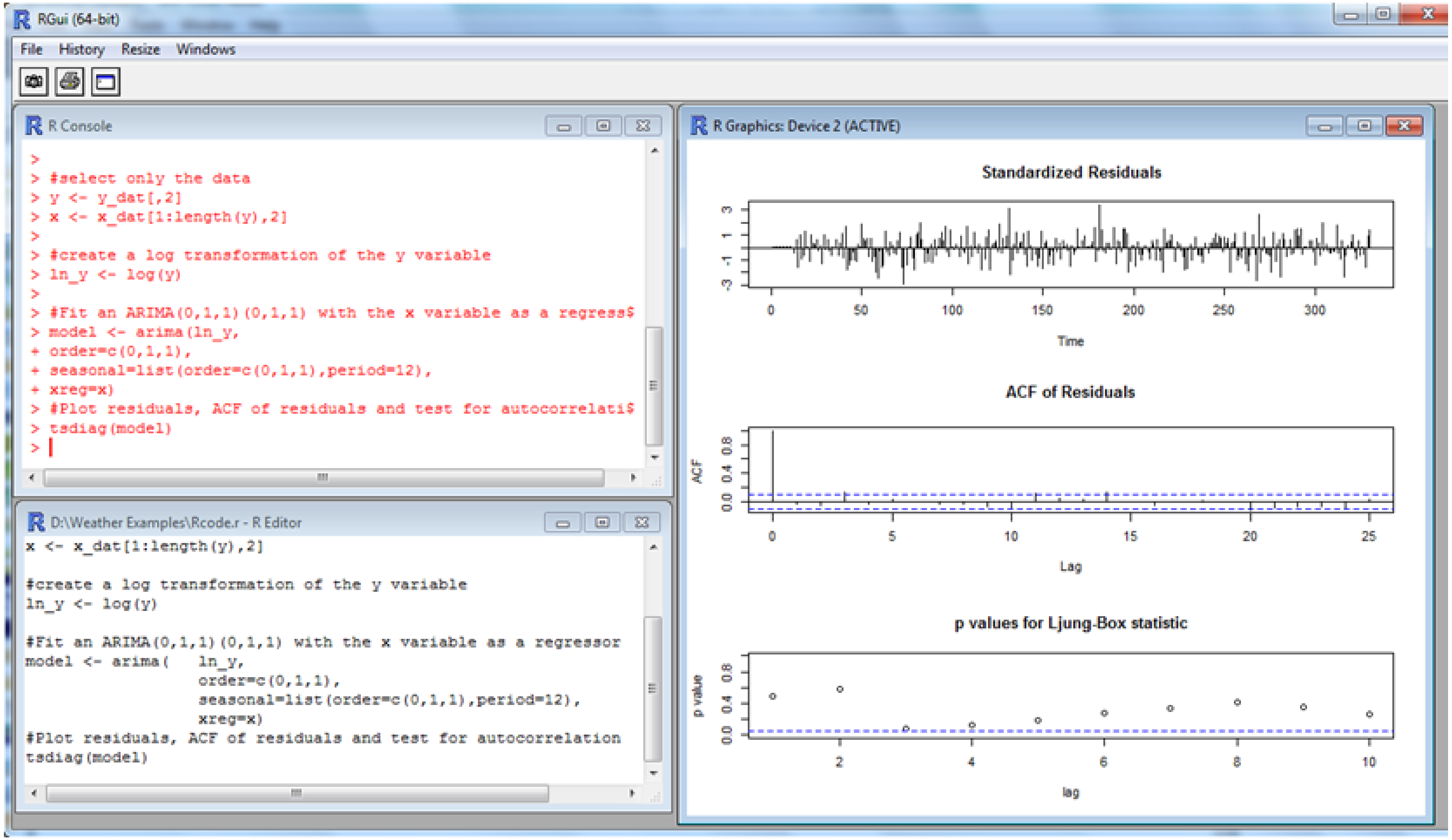
4.3.2 R

R is a freely available language and environment for statistical computing. In the following examples we use R version 3.0.2 for 64-bit windows (R Core Team, 2013). The code provided below is an example of fitting the same model described in the section on X-13ARIMA-SEATS. The code shows how to read in the y and x data that has been saved in the "x12save" format and fit the model used in the X-13ARIMA-SEATS example. As with X-13ARIMA-SEATS any lines of code that start with a "#" are comments and will not be executed.

```
#read in the x and y data
y_dat <- read.table(``D:\\Weather Examples\\Example 1\\y_data.dat'',skip=2)
x_dat <- read.table(``D:\\Weather Examples\\Example 1\\x_data.rmx'',skip=2)
#select only the data
y <- y_dat[,2]
x <- x_dat[1:length(y),2]
#create a log transformation of the y variable
ln_y <- log(y)
#Fit an ARIMA(0,1,1)(0,1,1) with the x variable as a regressor
model <- arima( ln_y,
                order=c(0,1,1),
                seasonal=list(order=c(0,1,1),period=12),
                xreg=x)
#Plot residuals, ACF of residuals and test for autocorrelation
tsdiag(model)
```

Back to flow chart

Figure 9: *Screenshot of R Gui showing plots from the tsdiag() function*



```
>
> #select only the data
> y <- y_dat[,2]
> x <- x_dat[1:length(y),2]
>
> #create a log transformation of the y variable
> ln_y <- log(y)
>
> #Fit an ARIMA(0,1,1)(0,1,1) with the x variable as a regress$
> model <- arima(ln_y,
+ order=c(0,1,1),
+ seasonal=list(order=c(0,1,1),period=12),
+ xreg=x)
> #Plot residuals, ACF of residuals and test for autocorrelati$
> tsdiag(model)
>
```

```
x <- x_dat[1:length(y),2]

#create a log transformation of the y variable
ln_y <- log(y)

#Fit an ARIMA(0,1,1)(0,1,1) with the x variable as a regressor
model <- arima(   ln_y,
                  order=c(0,1,1),
                  seasonal=list(order=c(0,1,1),period=12),
                  xreg=x)
#Plot residuals, ACF of residuals and test for autocorrelation
tsdiag(model)
```

Figure 9 shows a screen shot of the R Gui from executing the above lines of code. There are a couple of points worth noting. The x data from the rmx file started at the same date but ended at a later date than the y data. In order to use the arima() function in R the x and y data must have the same length if both are vector or the length of vector y must equal the number of rows in matrix x if more than one regressor is included in the model. The object “model” contains all the results from the model saved as a list object. The tsdiag() function provides plots of the standardised residuals, ACF and the p-values from the Ljung-Box statistics, which tests for autocorrelation in the residuals. For example, if the p-values are below 0.05 then this is an indication of autocorrelation. All the functions used in the above code are available from a basic installation of R. There are a number of contributed packages for time series analysis that are also available.

4.3.3 SAS

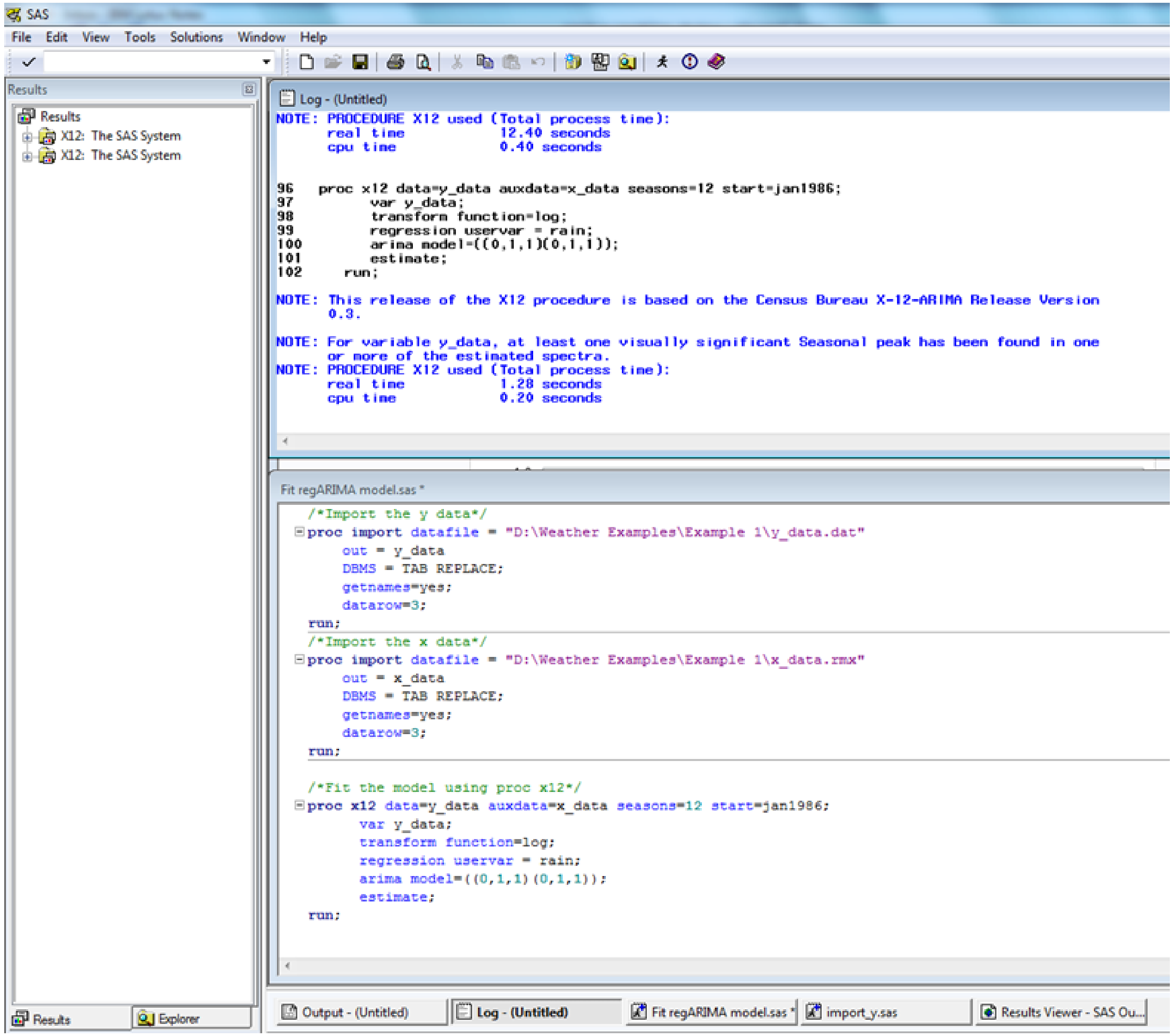
SAS® is commercially available software; the version used for the examples in this guide is SAS 9.3. The code below is used to replicate the analyses from sections 4.3.1 and 4.3.2. For fitting these models we use proc x12 from SAS/ETS®. Note that there are alternative procs available for fitting such models. In the example below we import the text files that were used in section 4.3.1. We leave the x and y data as separate datasets, although you could include them as one dataset. The output produced is very similar to that produced in the X-13ARIMA-SEATS output file.

```
/*Import the y data*/
proc import datafile = ''D:\Weather Examples\Example 1\y_data.dat''
       out = y_data
       DBMS = TAB REPLACE;
       getnames=yes;
       datarow=3;
run;

/*Import the x data*/
proc import datafile = ''D:\Weather Examples\Example 1\x_data.dat''
       out = x_data
       DBMS = TAB REPLACE;
       getnames=yes;
       datarow=3;
run;

/*Fit the model using proc x12*/
x12 data = y_data auxdata = x_data seasons = 12 start = jan1986
       var = y_data;
       transform function = log;
       regression uservar= rain;
       arima model=((0,1,1)(0,1,1));
       estimate;
run;
```

Figure 10 shows a screen shot after submitting the above code.

Back to flow chart

Figure 10: *Screenshot of SAS/ETS use of proc x12*



```
NOTE: PROCEDURE X12 used (Total process time):
      real time           12.40 seconds
      cpu time            0.40 seconds


     proc x12 data=y_data auxdata=x_data seasons=12 start=jan1986;
           var y_data;
           transform function=log;
           regression uservar = rain;
           arima model=((0,1,1)(0,1,1));
           estimate;
        run;

NOTE: This release of the X12 procedure is based on the Census Bureau X-12-ARIMA Release Version
      0.3.

NOTE: For variable y_data, at least one visually significant Seasonal peak has been found in one
      or more of the estimated spectra.
NOTE: PROCEDURE X12 used (Total process time):
      real time           1.28 seconds
      cpu time            0.20 seconds
```

```
/*Import the y data*/
proc import datafile = "D:\Weather Examples\Example 1\y_data.dat"
    out = y_data
    DBMS = TAB REPLACE;
    getnames=yes;
    datarow=3;
run;
/*Import the x data*/
proc import datafile = "D:\Weather Examples\Example 1\x_data.rmx"
    out = x_data
    DBMS = TAB REPLACE;
    getnames=yes;
    datarow=3;
run;

/*Fit the model using proc x12*/
proc x12 data=y_data auxdata=x_data seasons=12 start=jan1986;
      var y_data;
      transform function=log;
      regression uservar = rain;
      arima model=((0,1,1)(0,1,1));
      estimate;
run;
```

5 Communication

After completing any analysis you will need to consider what the results mean and how best to communicate these to your users. This section provides some guidance on interpreting and presenting your results. You are also strongly advised to consult the GSS guidance on writing about statistics.

Back to flow chart

5.1 Interpreting your results

If you have completed your analysis and found no suitable model that includes weather or climate variables, this cannot be interpreted as meaning there is no impact of weather or climate on your variable of interest. You can say that you have not found any evidence of a relationship between weather or climate data and your data of interest assuming that they are related in the way in which you have specified your model. If you find a model where a weather or climate variable appears to be statistically significant, your conclusions are based on the assumption that your model is correctly specified. These are important points to consider especially when communicating the results of any analysis. As discussed in the GSS guidance on writing about statistics care should be taken to use appropriate language.

Back to flow chart

5.2 Presenting results to users

When communicating your results it is necessary to consider the users of the data. Depending on the nature of the release, and perhaps the results of your analysis there are likely to be two options for communicating your results to a wider audience:

- Inclusion of the results in a regular publication
- A one-off article reporting your analysis

These options are briefly discussed in this section. If you have any questions or would like further support on the communication of your results contact the GSS Good Practice Team.

Including results in a regular publication

There are a number of ways in which results could be included in a regular publication. Here we discuss three of the more likely options:

- Publish a weather or climate adjusted time series.
- Publish seasonally adjusted time series where weather or climate is treated as a temporary prior adjustment.

- Do not adjust your time series for weather or climate but refer to your results in the commentary on your data.

The *ESS Guidelines on Seasonal Adjustment* (Eurostat, 2009) recommend **not** making weather adjustments to published official statistics. Such adjustments may be removing information from the time series that users are interested in. However, for some users and some publications there may be a case for making such adjustments. For example, DECC in its publication of energy statistics makes weather adjustments. The reason for this is that users are well aware of the relationship between weather and energy consumption and are not so interested in changes in energy consumption due to the climate and weather for short term analysis from regular publications.

If your time series is seasonally adjusted, there may already be some form of climate adjustment that is implicitly included in the seasonal factor. Note that this can only be interpreted as an average weather or climate effect for that period and it may also include other non-weather related seasonal effects.

When seasonally adjusting data, we sometimes refer to permanent or temporary prior adjustments to a time series.

A permanent adjustment is made for calendar or seasonal effects such as Easter or trading day effects. A temporary adjustment is made for outliers, level shifts or other types of effects not defined as seasonal. Both types of adjustments are made to improve the estimation of the seasonal component. The permanent adjustment ensures that the seasonally adjusted series does not have any systematic calendar related effects. The temporary adjustment improves the estimation of the seasonal component but is then added back in.

For example, if there is an outlier in a time series that has been caused by heavy snow in that period the effect is estimated and removed from the time series so as not to distort the estimation of the seasonal component, but is then shown in the seasonally adjusted time series.

If you can demonstrate that making a weather adjustment makes a substantial improvement to the estimation of your seasonal component then you may wish to consider regular inclusion of the weather or climate adjustment as a temporary prior adjustment where you publish seasonally adjusted data. Note that this should be regularly reviewed to ensure that the model is appropriate.

Writing a one-off article

It is recommended to write an article to describe in detail the methods and results of your analysis. It may also be useful for users for you to publish a non-technical article describing the main conclusions of your analysis in a way that is accessible to those who are not familiar with the technical details. These articles may then be referred to in the commentary in regular publications where appropriate. If the results of your analysis are only of general interest to users of the data, or if you intend to use some of the results in a regular publication, a one-off article will be useful as a point of reference.

Back to flow chart

6 Retail Sales

6.1 Overview

This case study looks at the effect of temperature on retail sales for clothing. The series used in the analysis were the retail sales of clothing series from the ONS and the monthly mean UK temperature series, one of the monthly climate series from the Met Office. Initial graphical analysis and discussions with the statistician responsible for retail sales statistics led to the testing of a switching effect where clothing sales are brought forward in warmer than average springs and pushed back in warmer than average autumns. regARIMA modelling was used to test and estimate this effect. The model found significant positive effects (bringing purchases forward if the month is warmer than average) in all months from February to July, and negative switching effects in September, October and November. The largest effect was found in May where a 1C rise in temperature above average increases the level of sales in May by 1.6% and reduces the level of sales in June by 1.6%.

6.2 Clarify objectives

To understand how weather or climate affects retail sales of clothing.

6.3 Plan the investigation and select appropriate data

6.3.1 Plan the investigation

The plan for the investigation was as follows:

1. Collect and understand the data available on retail sales of clothing.

2. Choose an appropriate source of weather or climate data.

3. Carry out preliminary analysis to identify potential weather effects.

4. Discuss potential weather effects with statisticians working on retail sales statistics and formulate a hypothesis to test.

5. Create appropriate regressors to include in regARIMA modelling.

6. Define a regARIMA model with a full set of regressors to test.

7. Run backwards selection.

8. Analyse model diagnostics and assess stability.

9. Repeat regARIMA modelling as necessary based on results from each iteration.

10. Interpret the results.

Back to flow chart

6.3.2 Select appropriate data

Retail sales data

The retail sales index measures changes in value and volume of sales of retail goods and is produced by ONS. As well as the overall index, series are available for specific parts of the retail sector, clothing being one of them. The series that was selected for this analysis was the current price value series of clothing sales.

What does the series measure? Changes value of sales of clothing in current prices.

How frequent is the series? Monthly.

How long is the series? January 1986 to June 2014.

Met Office data

The weather data selected for this analysis were from the monthly climate series.

What does the series measure? Mean monthly UK temperature.

How frequent is the series? Monthly.

How long is the series? January 1910 to June 2014.

Back to flow chart

6.4 Assess the data quality and structure

6.4.1 Retail sales data

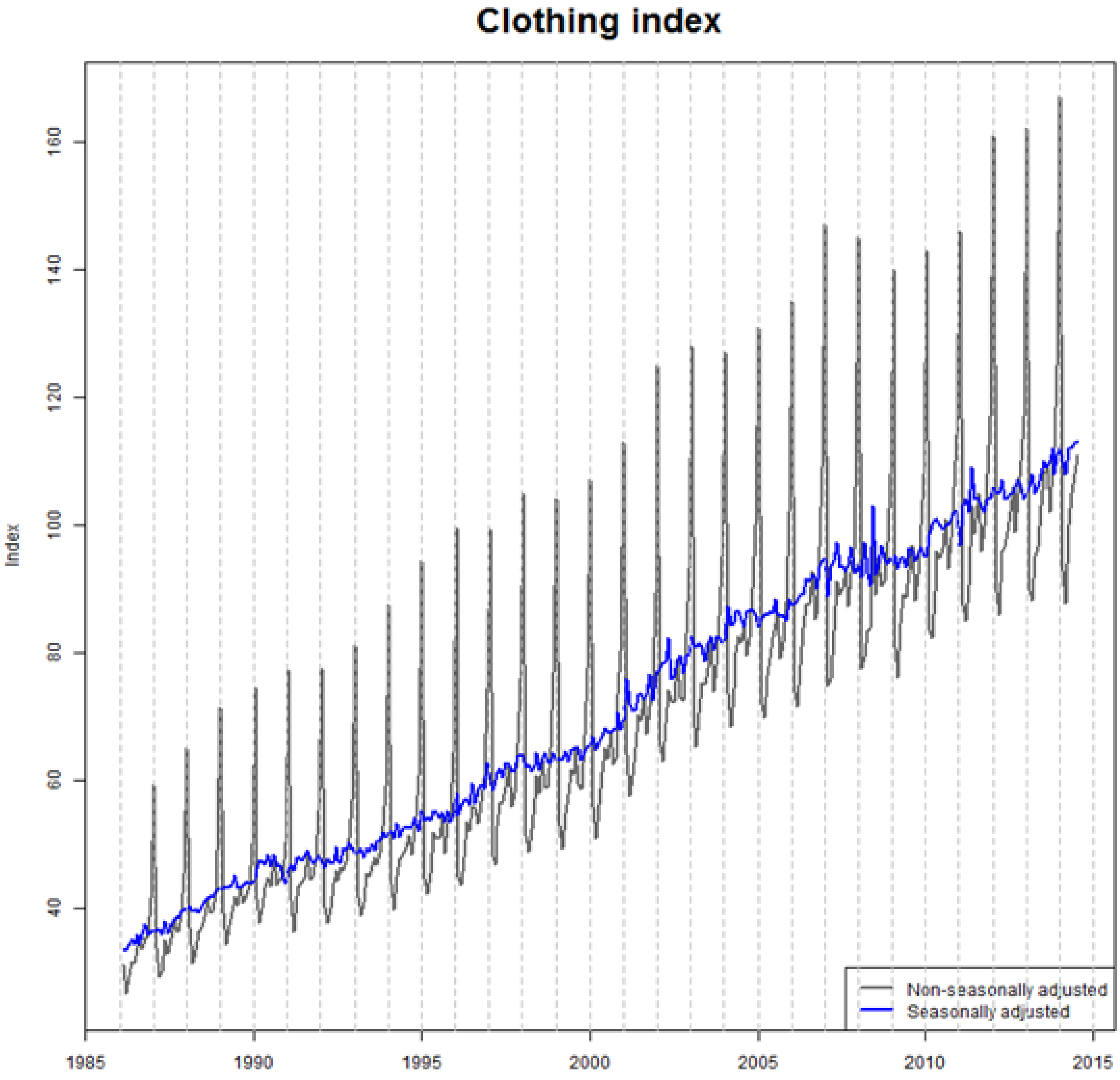
How is the series produced? Total retail turnover data are collected on a monthly basis from retailers. Rather than reporting turnover for the calendar month, businesses are asked for their turnover in either a 4 or 5 week period. These data are used to estimate average weekly turnover for the month. These are then converted into an index. More information on how retail sales estimates are compiled is available in "A Quick Guide to the Retail Sales Index".

What does the series look like? Both seasonally adjusted and non-seasonally adjusted clothing indices are plotted in figure 11. The clothing series is highly seasonal, with regular peaks every December. Both series exhibit an upwards trend. Deviations from the trend that are not caused by seasonal factors (known as the irregular component of a time series) can be seen in the seasonally adjusted series. These deviations are potentially of interest in terms of weather analysis as they could be due to weather effects.

Seasonal adjustment of retail sales series. Retail sales series are seasonally adjusted. Series are seasonally adjusted for two reasons: to remove effects due to the time of year and to remove effects due to the arrangement of the calendar. Retail sales of clothing contain effects due to the time of year as figure 11 clearly shows that the index is highest each year in December. The calendar effects come from the fact that the reporting periods do not line up with calendar months, as well as holidays for example Easter moving between reporting periods. To account for this, regARIMA modelling is used prior to seasonal adjustment which looks at the difference between the centre of the reference month and the reporting period as well as where moving holidays that affect the series fall.

Back to flow chart

Figure 11: *Retail sales of clothing seasonally adjusted and non-seasonally adjusted from January 1986 to June 2014.*



6.4.2 Met Office data

How is the series produced? See the guide to weather and climate data.

What does the series look like? Figure 12 plots the mean monthly UK temperature series from 1986. It can be seen in figure 12 that, as expected, UK mean monthly temperature is seasonal with colder temperatures in the winter months and warmer temperatures in the summer months. From this type of plot, colder than usual winters and warm than usual summers can be identified, but its more difficult to see what is happening in other months.

Visualising extreme weather. In order to visualise whether a particular month was warmer or colder than usual, the deviations from the long running monthly average temperature were calculated. These were calculated separately for each month, but in the same way, so the January one was calculated by taking the mean of all of the January temperatures between January 1986 and July 2014, then subtracting this from each of the January temperatures. These series can be seen in figure 13. The separate monthly deviation series can be combined to form a time series of all months from January 1986 to July 2014. This series is shown in figure 14.

Back to flow chart

Figure 12: *Mean monthly UK temperature from January 1986 to June 2014.*

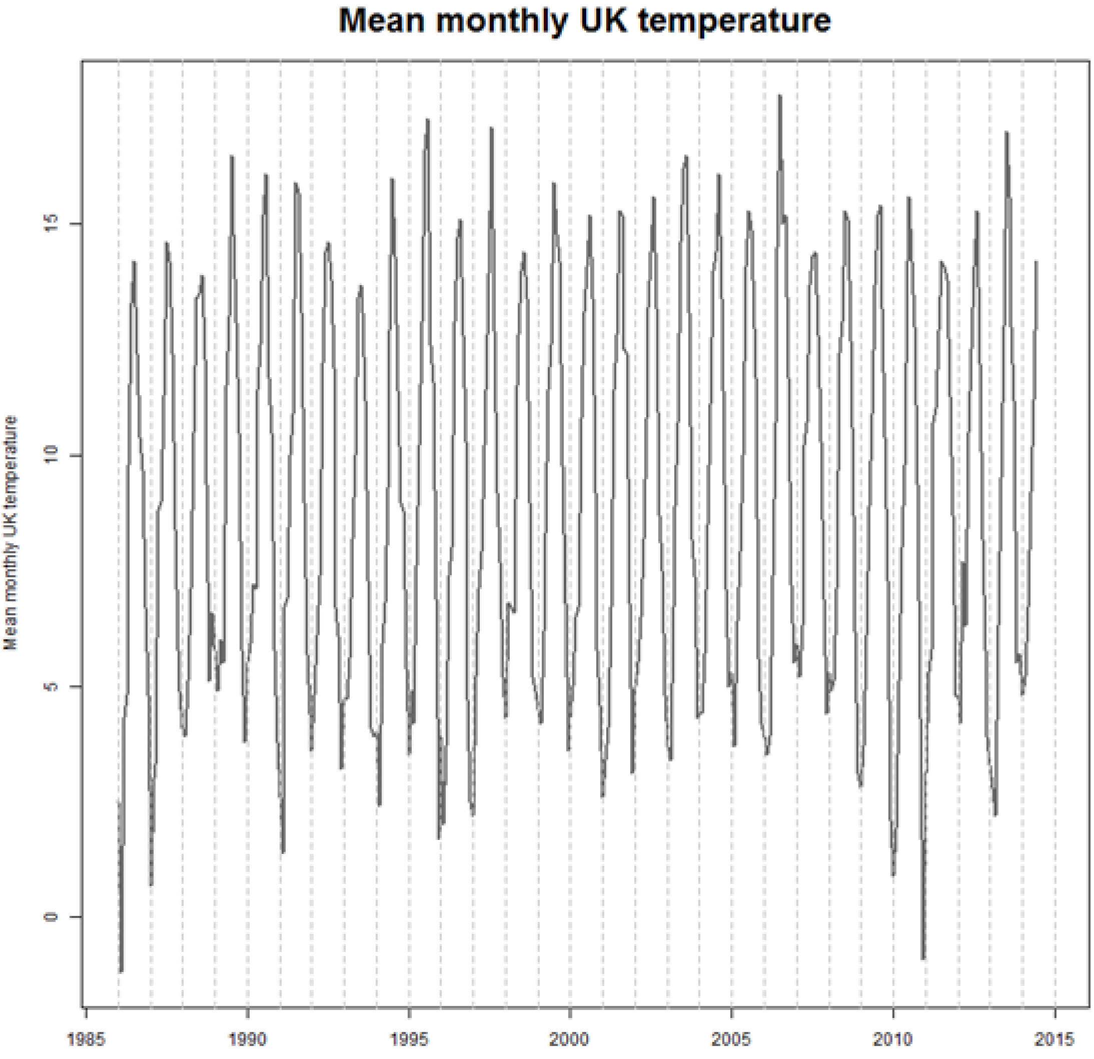


Figure 13: *Temperature deviation from long-run monthly average calculated between January 1986 and June 2014.*

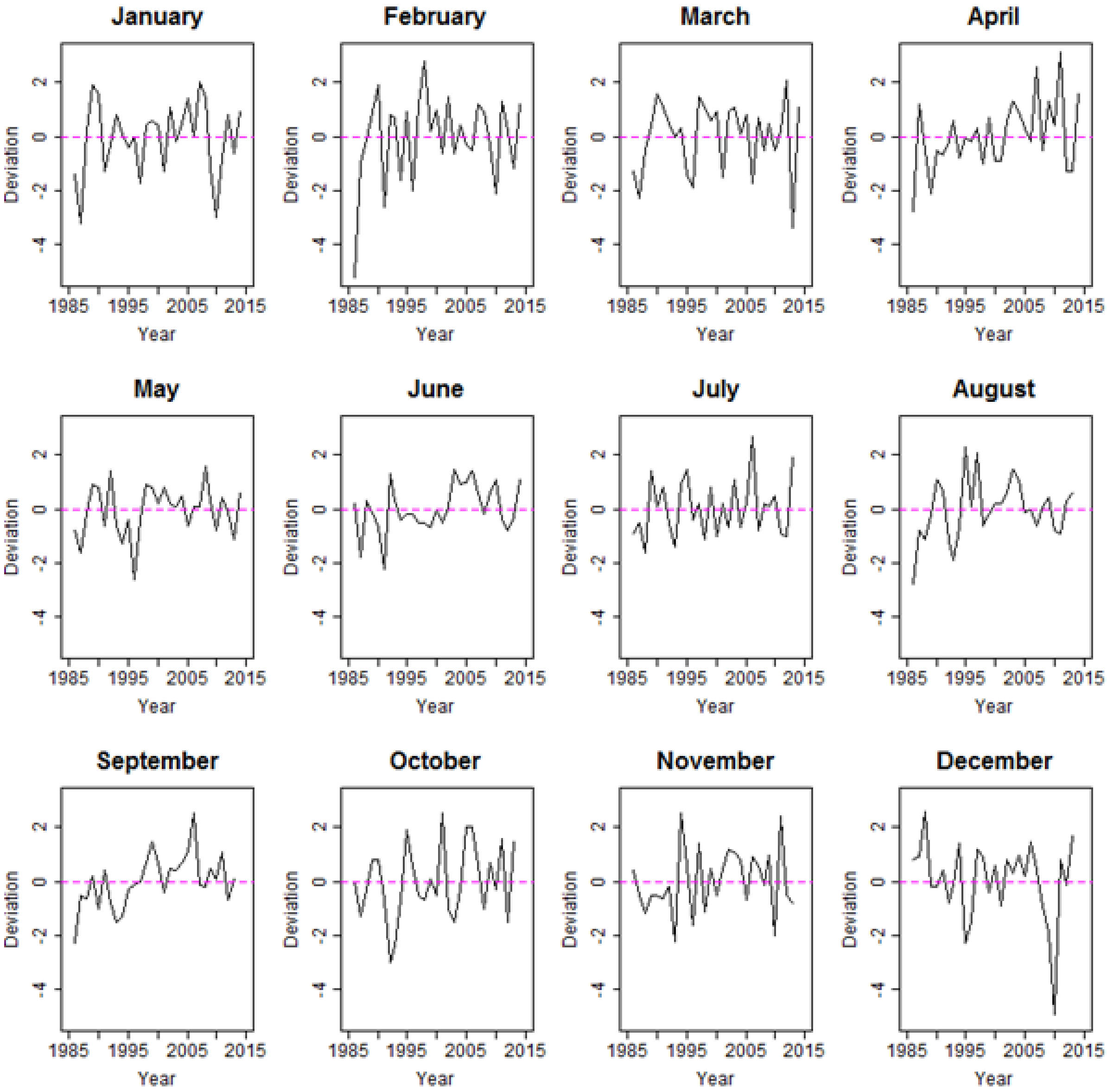
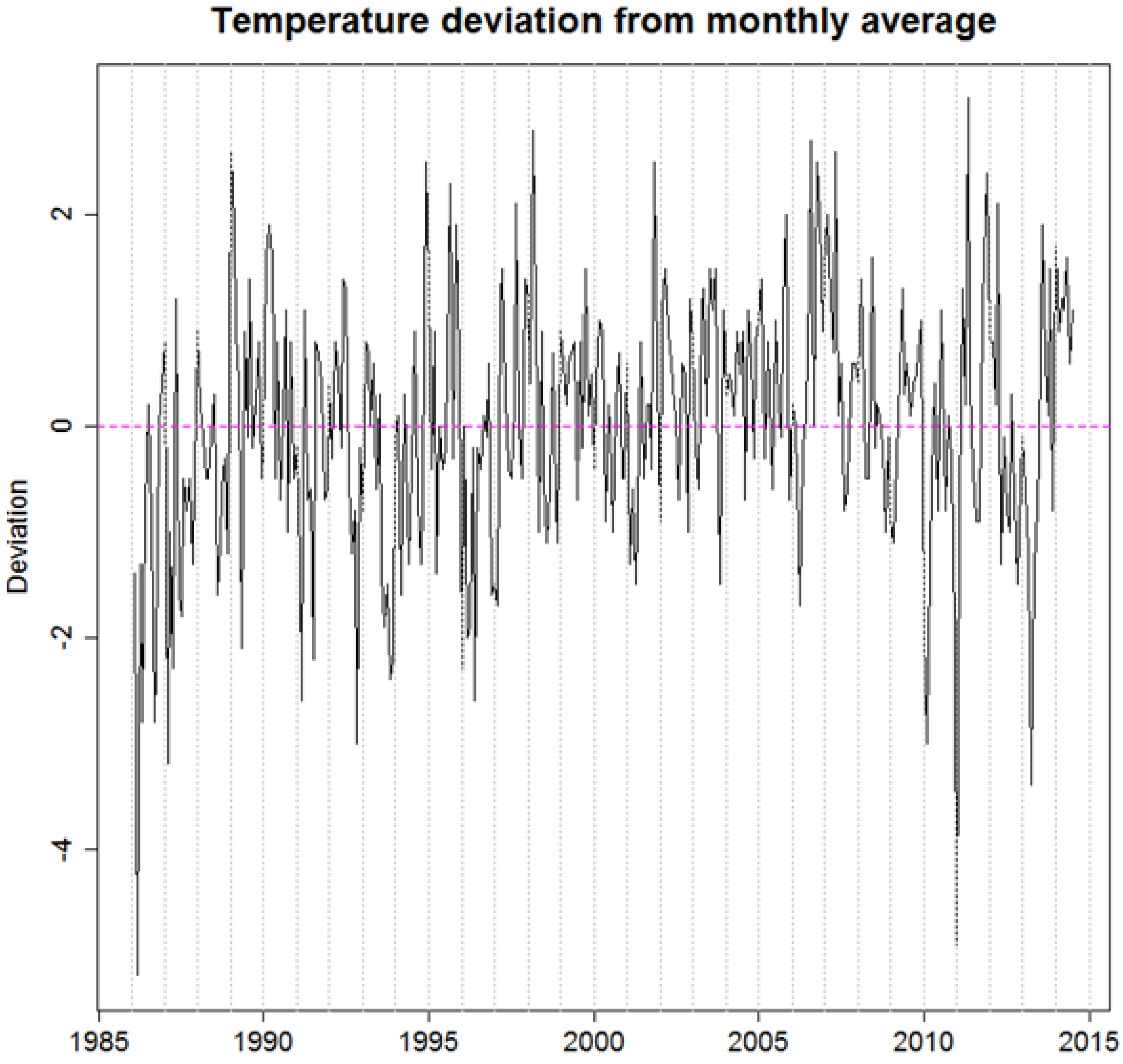


Figure 14: *Time series of temperature deviation from long-run monthly average between January 1986 and June 2014.*



6.5 Initial data analysis

6.5.1 Visualise the clothing and weather data

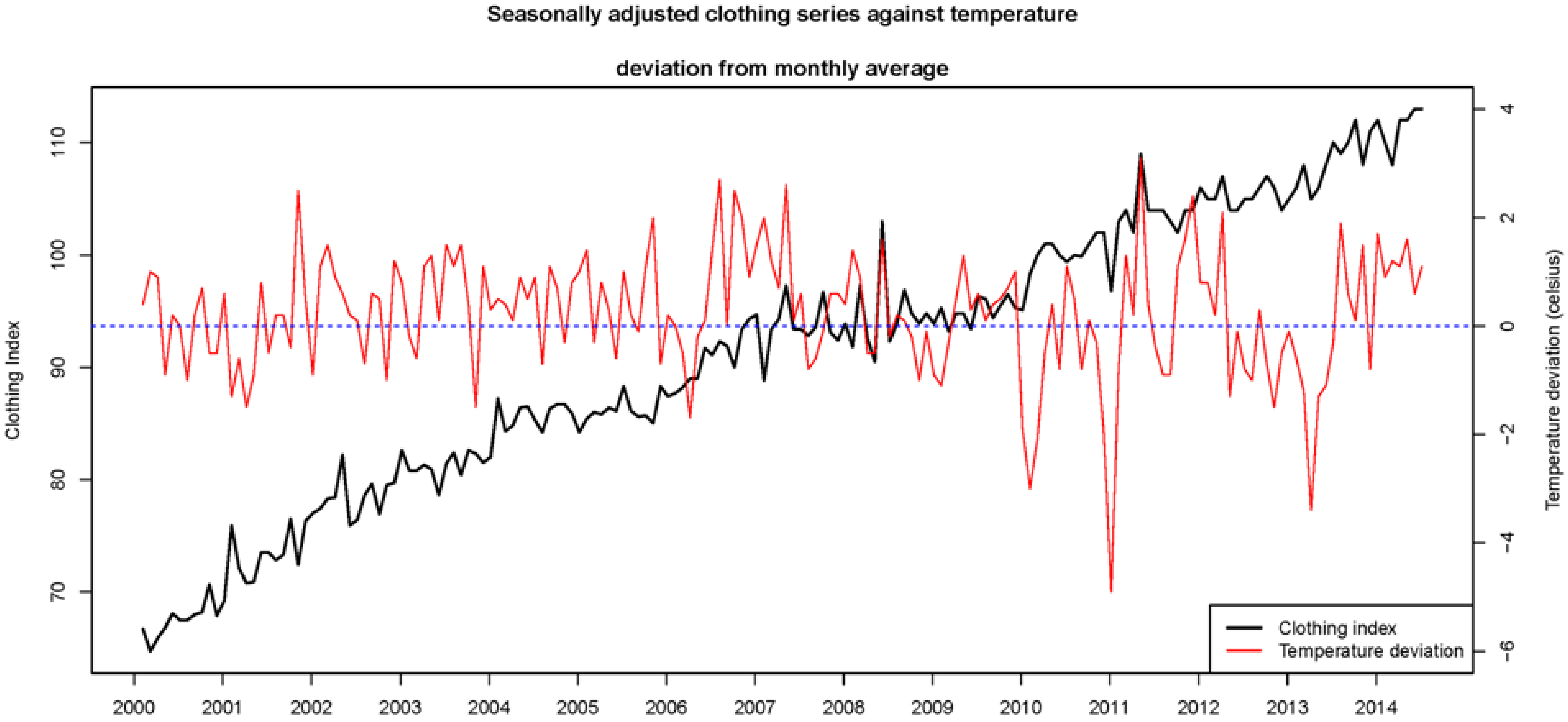
Initial data analysis focussed on visualising the clothing and weather series together to allow visual identification of any potential relationships. The seasonally adjusted clothing series and the temperature deviations from the long-run monthly average series have been plotted on one graph to make it easy to see when extremes in temperature coincide with dips or peaks in the clothing series. This is shown in figure 15.

The seasonally adjusted series has been used to allow easier visualisation of deviations from the trend. Deviations in temperature are used to allow identification of warmer or colder than average months.

From the graph, it can be seen that there are periods of colder or warmer weather which coincide with the visually noticeable peaks or troughs in the seasonally adjusted series. For example, the cold weather in December 2010 and March 2013 correspond to dips in clothing sales. If there is an effect then it may vary between months as in October 2001 and January 2007 the temperature was warmer than average and clothing sales dipped, while in April 2011, the temperature was warmer than average and clothing sales increased.

This sort of analysis has already been used within the GSS, for example an ONS Retail Sales article, titled ‘How sensitive to the weather is the retail sector?’ (ONS, 2014).

Figure 15: *Seasonally adjusted clothing series against temperature deviation from long-run monthly average.*

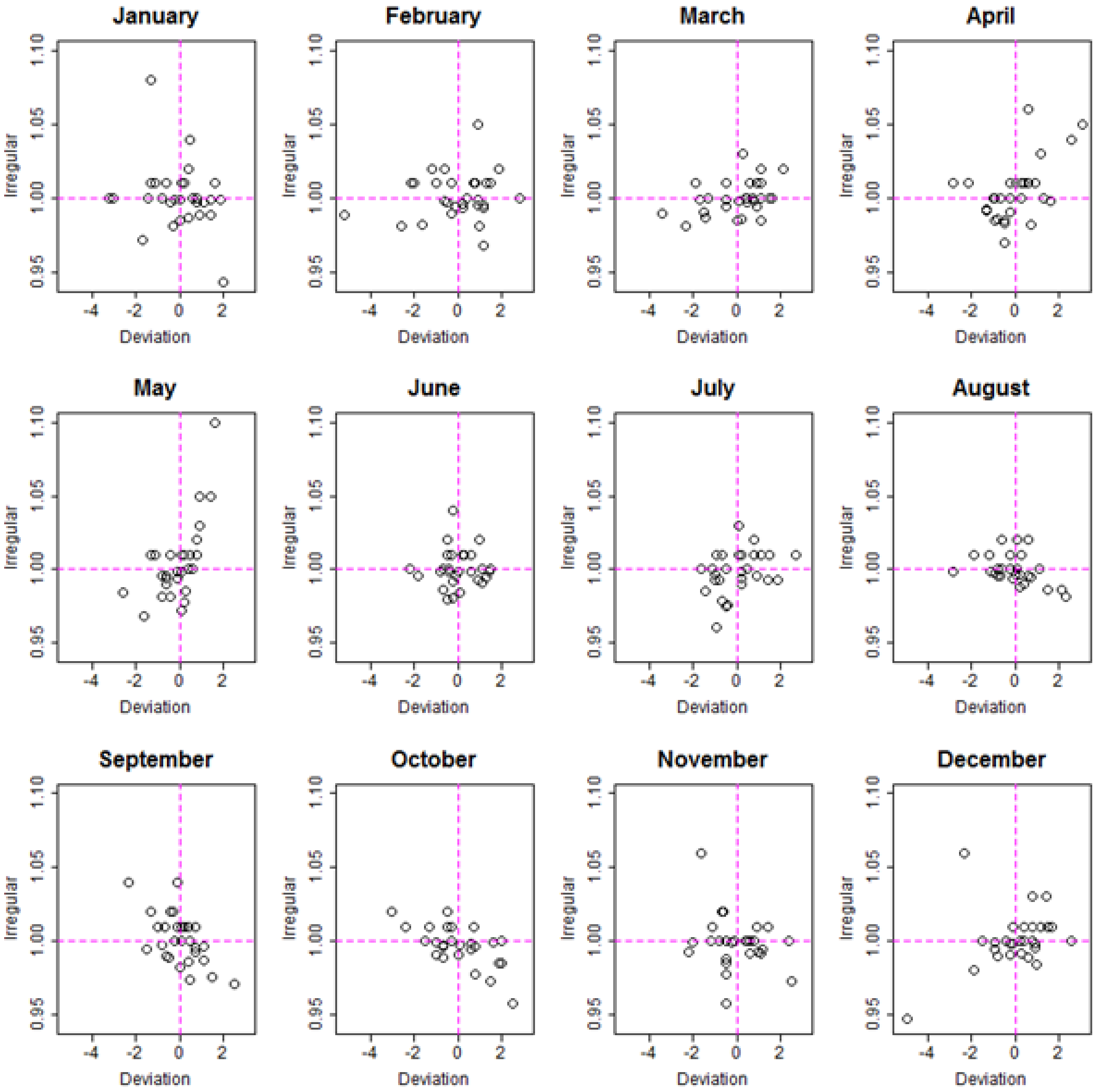


The irregular component of the seasonally adjusted series contains variations within a time series that are not explained by the long running trend or the seasonality. If the weather does affect clothing sales then the effect might be visible in the irregular component of the time series.

The irregular component of the seasonally adjusted series (from seasonal adjustment using the X-11 algorithm in X-13ARIMA-SEATS) was plotted against the temperature deviation separately for each month to see whether there were any visually noticeable relationships between higher or lower unexplained variation in the time series and higher or lower than average temperature. This was carried out separately for each month to allow for different effects in different months. The results of this can be seen in figure 16. These charts indicate a positive correlation between unexplained movements in clothing sales and deviations from average temperature for March, April and May and a negative correlation for September and October. In the other months such relationships are less clear.

Back to flow chart

Figure 16: *Irregular component of decomposed clothing time series against temperature deviation from long-run monthly average.*



6.5.2 Discuss potential weather effects and formulate a hypothesis to test

Initial investigations of the retail data through graphical analysis suggested that there might be weather effects in the spring and autumn where purchases of summer clothes might be brought forward in spring if it is warmer than usual and purchases of winter clothes pushed back in autumn if it is warmer than usual. Given the results of the graphical analysis and based on discussions with the statistician responsible for retail sales data, regARIMA modelling was used to test for a switching effect in sales of clothes between months if it was warmer or colder than usual.

Back to flow chart

6.6 Time series modelling

6.6.1 Create appropriate regressors to include in regARIMA modelling

The weather regressors were calculated by taking the deviation in temperature in a given month from the average temperature in that month over the period January 1986 to June 2014. The regressor for a particular month took the deviation if the time period was that month, the negative deviation of the previous month if the time period is the month after, and 0 otherwise. This is illustrated in equation 14. Phase shift regressors were also included.

$$x_{i,t} = \begin{cases} tempdev_t & \text{if } month(t) = i \\ -tempdev_{t-1} & \text{if } month(t-1) = i \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Where $tempdev_t$ is the temperature deviation from the long run average at time-point t.

Back to flow chart

6.6.2 Define a regARIMA model with a full set of regressors to test

The retail sales series shows a clear multiplicative relationship between components of the series over time and so a log transformation of the series is used. The automatic selection of the order of the ARIMA model in X-13ARIMA-SEATS was an $(0,1,2)(0,1,1)_{12}$ regARIMA model. Results are presented below.

6.6.3 Run backwards selection

The model was run initially with all regressors included. Backwards selection was then used to select a model with a set of regressors which were all statistically significant. Backwards selection is an iterative process in which all of the regressors that are not found to be statistically significant, the least significant is removed. The model is then refitted and the process is run again until all remaining regressors are significant.

The resulting regressors in the final model and their estimated coefficients can be found in the examples of output provided below. The "temp_mmm" variables are the temperature deviations for the month "mmm". The Nov and Dec variables are phase shift effects for November and December (other months were not found to be significant), and Eas1 is an Easter regressor.

The model has been fitted in Win X-13, R and SAS. The files required to produce these outputs is also provided. *Note: The estimated coefficients vary slightly between the different softwares due to differences in the estimation methods. They may also vary between different versions of the same software.*

regARIMA model output from Win X-13.

```
Regression Model
------------------------------------------------
                      Parameter  Standard
Variable               Estimate     Error    t-value
------------------------------------------------
User-defined
temp_feb                 0.0044   0.00138       3.15
temp_mar                 0.0044   0.00166       2.63
temp_apr                 0.0109   0.00176       6.19
temp_may                 0.0161   0.00243       6.61
temp_jun                 0.0065   0.00266       2.44
temp_jul                 0.0049   0.00202       2.43
temp_sep                -0.0130   0.00279      -4.65
temp_oct                -0.0100   0.00179      -5.56
temp_nov                -0.0076   0.00170      -4.44
Nov                     -0.0096   0.00199      -4.83
Dec                      0.0054   0.00184       2.93
Eas1                     0.0114   0.00489       2.34
------------------------------------------------
```

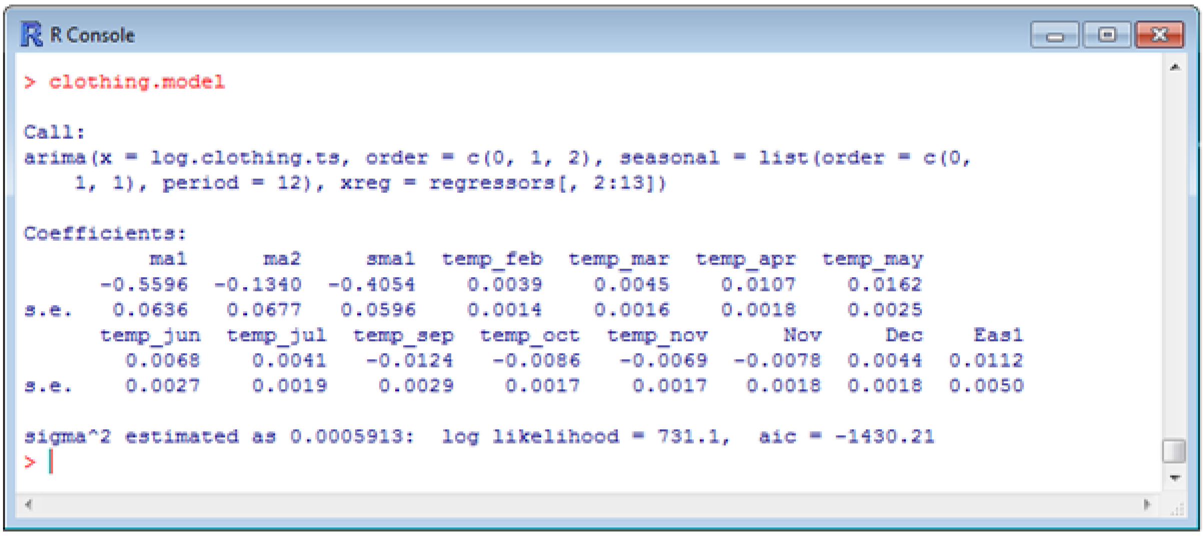
Back to flow chart

6.6.4 Data and code

Files required to carry out analysis in Win X-13

Right click here to save spec file
Right click here to save data file
Right click here to save regressors file

regARIMA model output from R.

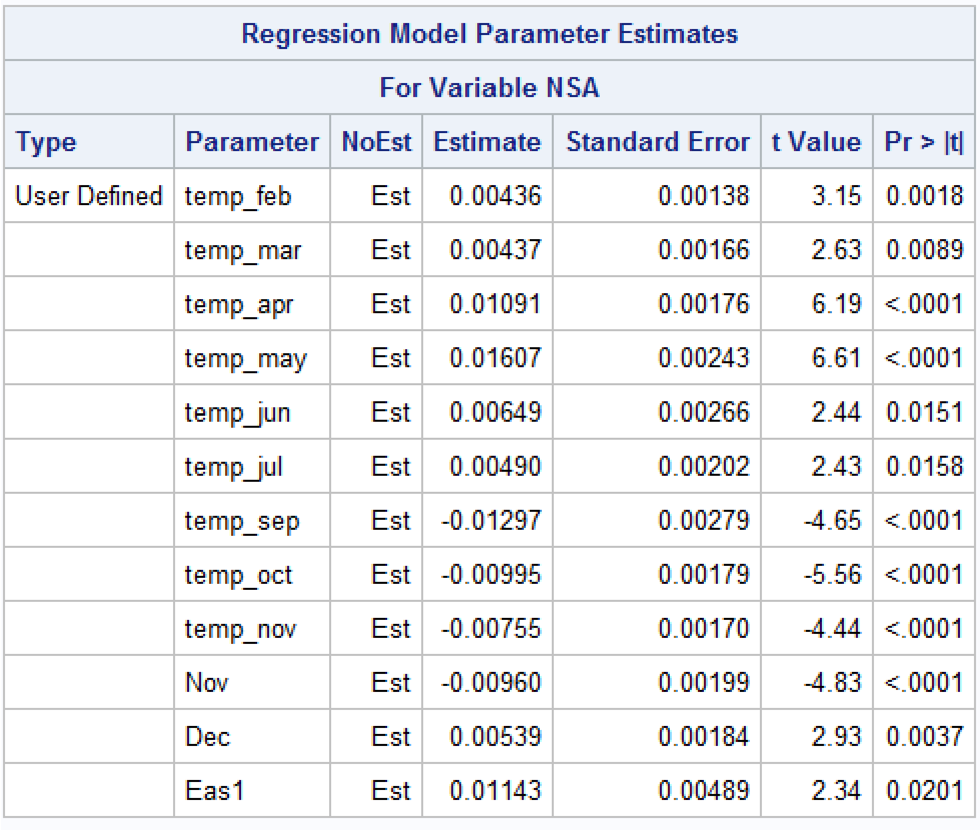


```
> clothing.model

Call:
arima(x = log.clothing.ts, order = c(0, 1, 2), seasonal = list(order = c(0,
    1, 1), period = 12), xreg = regressors[, 2:13])

Coefficients:
          ma1      ma2     sma1  temp_feb  temp_mar  temp_apr  temp_may
      -0.5596  -0.1340  -0.4054    0.0039    0.0045    0.0107    0.0162
s.e.   0.0636   0.0677   0.0596    0.0014    0.0016    0.0018    0.0025
      temp_jun  temp_jul  temp_sep  temp_oct  temp_nov      Nov     Dec    Eas1
        0.0068    0.0041   -0.0124   -0.0086   -0.0069  -0.0078  0.0044  0.0112
s.e.    0.0027    0.0019    0.0029    0.0017    0.0017   0.0018  0.0018  0.0050

sigma^2 estimated as 0.0005913:  log likelihood = 731.1,  aic = -1430.21
> 
```

Files required to carry out analysis in R

Right click here to save R code file
Right click here to save data file
Right click here to save regressors file

regARIMA model output from SAS.



Regression Model Parameter Estimates						
For Variable NSA						
Type	Parameter	NoEst	Estimate	Standard Error	t Value	Pr > |t|
User Defined	temp_feb	Est	0.00436	0.00138	3.15	0.0018
	temp_mar	Est	0.00437	0.00166	2.63	0.0089
	temp_apr	Est	0.01091	0.00176	6.19	<.0001
	temp_may	Est	0.01607	0.00243	6.61	<.0001
	temp_jun	Est	0.00649	0.00266	2.44	0.0151
	temp_jul	Est	0.00490	0.00202	2.43	0.0158
	temp_sep	Est	-0.01297	0.00279	-4.65	<.0001
	temp_oct	Est	-0.00995	0.00179	-5.56	<.0001
	temp_nov	Est	-0.00755	0.00170	-4.44	<.0001
	Nov	Est	-0.00960	0.00199	-4.83	<.0001
	Dec	Est	0.00539	0.00184	2.93	0.0037
	Eas1	Est	0.01143	0.00489	2.34	0.0201

Right click here to save SAS code file
Right click here to save data file

Back to flow chart

6.6.5 Analyse model diagnostics and assess stability

Model diagnostics have been taken from the output when running the modelling in Win X-13.

Likelihood statistics

Win X-13 produces a set of likelihood statistics assessing the fit of the regARIMA model. The estimated likelihood statistics for the clothing model are provided below. AIC values can be compared for nested models (where all regressors included in one model are included in the other model). Thus, to see if the weather regressors improve the model fit, the AIC of the model including weather regressors can be compared to the AIC of the model including just the phase shift regressors that are in the final model (Nov, Dec and Eas1) or the model with no regressors. The likelihoods of these models are also provided below. The AIC test statistics (highlighted in bold and red) are lowest when the weather regressors are included in the model.

Likelihood statistics for model including weather regressors.

```
--------------------------------------------------------
Number of observations (nobs)                        330
Effective number of observations (nefobs)            317
Number of parameters estimated (np)                   16
Log likelihood                                  736.3777
Transformation Adjustment                     -1330.4939
Adjusted Log likelihood (L)                    -594.1162
AIC                                            1220.2325
AICC (F-corrected-AIC)                         1222.0458
Hannan Quinn                                   1244.2564
BIC                                            1280.3749
--------------------------------------------------------
```

Likelihood statistics for model including phase shift regressors.

```
--------------------------------------------------------
Number of observations (nobs)                        330
Effective number of observations (nefobs)            317
Number of parameters estimated (np)                    7
Log likelihood                                  677.2688
Transformation Adjustment                     -1330.4939
Adjusted Log likelihood (L)                    -653.2252
AIC                                            1320.4503
AICC (F-corrected-AIC)                         1320.8128
Hannan Quinn                                   1330.9608
BIC                                            1346.7627
--------------------------------------------------------
```

Likelihood statistics for model not including weather or phase shift regressors.

```
--------------------------------------------------------
Number of observations (nobs)                        330
Effective number of observations (nefobs)            317
Number of parameters estimated (np)                    4
Log likelihood                                  664.8812
Transformation Adjustment                     -1330.4939
Adjusted Log likelihood (L)                    -665.6127
AIC                                            1339.2254
AICC (F-corrected-AIC)                         1339.3536
Hannan Quinn                                   1345.2314
BIC                                            1354.2610
--------------------------------------------------------
```

Normality tests

To assess normality of the regARIMA model residuals, three test statistics were calculated; skewness, Gearys a and kurtosis. The results of these tests are provided below. None of these tests indicates a lack of normality in the model residuals. *Note: For a normal distribution, skewness is 0, Geary's a is* $\sqrt{\frac{2}{\pi}}$ *and kurtosis is 3.*

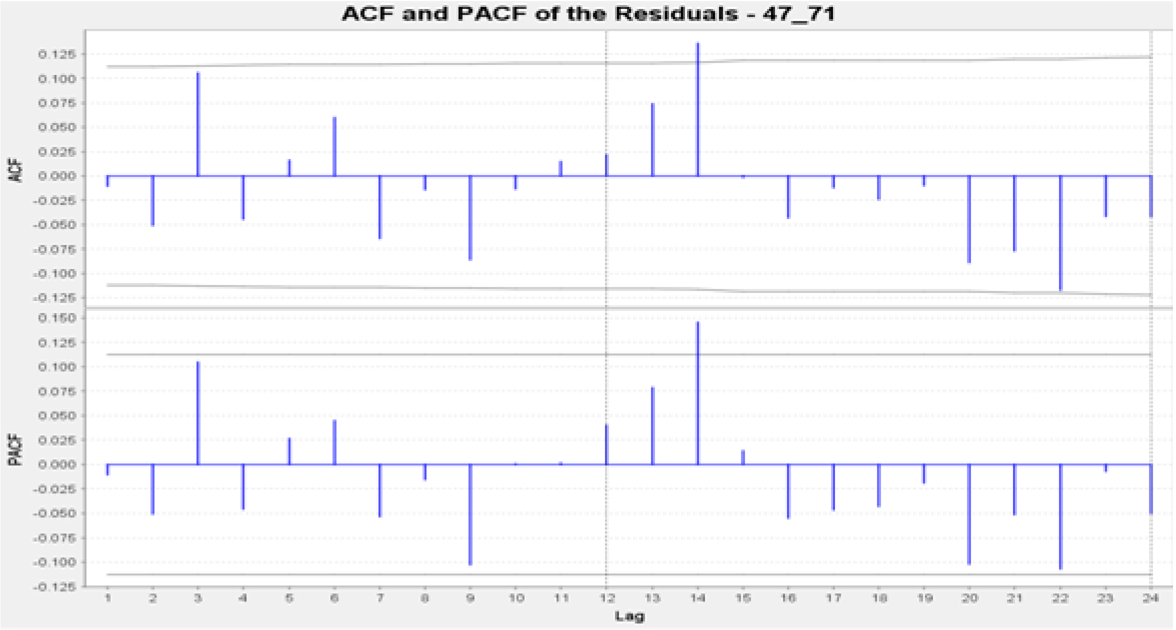
Normality Statistics for regARIMA Model Residuals.

```
Number of residuals     :  317
Skewness coefficient    :  0.0212
Geary’s a               :  0.7830
Kurtosis                :  3.6302

No indication of lack of normality
```

Autocorrelation and partial autocorrelation

Plots of the autocorrelation and partial autocorrelation functions of the model residuals are provided in figure 17. Both the autocorrelation and partial autocorrelation functions have significant peaks at lag 14, however there are no peaks at shorter lags.

Figure 17: *Auto correlation and partial autocorrelation function of the model residuals.*



Ljung-Box test statistics can also be used to assess whether significant autocorrelation exists in the model residuals. The output is provided below. As described in the summary provided by Win X-13 (below the test statistics), where the degrees if freedom (DF) are positive, P values below 0.05 may indicate model inadequacy. There is some indication at lags 4, 14 and 24.

Sample Autocorrelations of the Residuals with the Ljung-Box diagnostic.

```
Lag       1      2      3      4      5      6      7      8      9     10     11     12
ACF       0  -0.06    0.1  -0.03   0.01      0  -0.03  -0.04  -0.08  -0.07   0.11   0.06
SE     0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06
Q         0    1.3    4.8   5.17   5.23   5.23   5.59   6.25   8.38   9.89  13.79  14.87
DF        0      0      0      1      2      3      4      5      6      7      8      9
P         0      0      0  0.023  0.073  0.156  0.232  0.283  0.211  0.195  0.088  0.095

Lag      13     14     15     16     17     18     19     20     21     22     23     24
ACF    0.04   0.13      0  -0.05  -0.01   0.02      0   -0.1  -0.09  -0.06  -0.08   -0.1
SE     0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06
Q     15.45  20.79  20.79  21.58  21.65   21.8  21.81  25.51  28.34  29.71  31.72   35.3
DF       10     11     12     13     14     15     16     17     18     19     20     21
P     0.117  0.036  0.053  0.062  0.086  0.113   0.15  0.084  0.057  0.056  0.046  0.026

The P-values approximate the probability of observing a Q-value at least
this large when the model fitted is correct.  When DF is positive, small
values of P, customarily those below 0.05, indicate model inadequacy.
```

Trading day effects

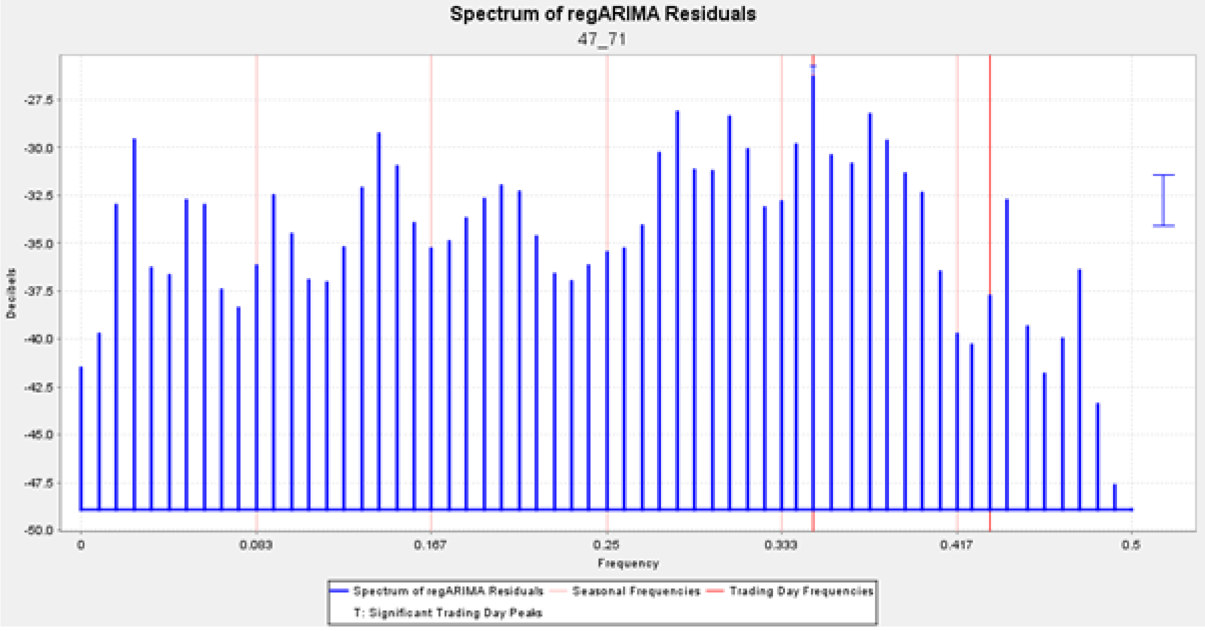
A spectral plot of the model residuals (figure 18) indicates a visually significant peak at a trading day frequency. Trading day regressors are not included in the retail sales seasonal adjustment because the data are collected on a 4-4-5 week basis, so that each reporting period will have the same number of each trading day.

Seasonality

Win X-13 includes the QS statistic which tests the hypothesis of no seasonality. This test is carried out for the regARIMA model residuals and the results are provided below. These tests show that there is no seasonality in the residuals.

```
QS Statistic for regARIMA Model Residuals (full series):                       0.18
                                                                  (P-Value = 0.9142)
QS Statistic for regARIMA Model Residuals (starting 2005.Jul):                 0.00
                                                                  (P-Value = 1.0000)
```

Figure 18: *Spectral plot of regARIMA model residuals.*



Analysis on different spans of data

Some work has been done to assess the stability of the regARIMA model, by fitting the model to different spans of the data. When using spans of data starting from January 1994, 1996, 1998, 2000 and 2002 to the end of the series, the February, March, June and July regressors became insignificant, leaving significant effects in April, May, September, October and November. The estimated coefficients for those months that remain significant in all spans of data remained relatively stable with an April effect of around 1.1% to 1.5%, May 1.6% to 2.2%, September -1.3% to -1.6%, October -1% to -1.1% and November -0.7% to -0.9%.

Back to flow chart

6.6.6 Repeat regARIMA modelling as necessary based on results from each iteration

Further work is required to assess the stability and fit of the model. This may lead to refinements to the model.

One potential refinement is in the construction of the regressors. The way that the switching effect has been constructed allows for switching between all pairs of consecutive months. However, it may be that summer clothes are purchased in the first month that the temperature reaches a certain point. For example, in the model, if March was warmer than usual and April colder, then April sales would be both brought forward to March and pushed back to May.

Back to flow chart

6.7 Interpretation and communication of the results

The significant regressors suggest that there is a positive switching effect in February, March, April, May, June and July (bringing purchases forward if the month is warmer than average) and a negative switching effect in September, October and November (pushing purchases back if the month is warmer than average). The largest effect is found for a May regressor with an estimated coefficient of 0.016 which implies that if the temperature in May is 1°C higher than average this will increase the level of sales in May by about 1.6% and reduce the level of sales in June by about 1.6%, all else being equal.

Back to flow chart

7 Road Accidents

7.1 Overview

Analysis of road accident data focused on the effect of temperature on monthly counts of killed or seriously injured road users. The road accident series was compiled from administrative records of all police recorded road accidents, published by the Department for Transport. The weather data used in this analysis were from the monthly mean UK temperature series, one of the monthly climate series from the Met Office. Initial analysis suggested that if there is an effect then the effect might vary between months. regARIMA modelling was used to test in which months there is an effect of temperature on the number of killed or seriously injured vulnerable road users, and also estimate the effect. The final model found significant positive effects (above average temperatures leading to increases, below average temperatures leading to decreases) for all months of the year except for August, September and November. The effect varies between months, from a 1°C increase in temperature above average (all else being equal) resulting in an additional 27 killed or seriously injured vulnerable road users in October to an additional 95 in April. Similarly, according to the model a 1°C decrease in temperature below average results in between 27 fewer killed or seriously injured vulnerable road users in October and 95 fewer in April.

Back to flow chart

7.2 Clarify objectives

To understand how weather or climate affects the number of killed or seriously injured vulnerable road users.

7.3 Plan the investigation and select appropriate data

7.3.1 Plan the investigation

The plan for the investigation was as follows.

1. Collect and understand the data available on road accidents.
2. Choose an appropriate source of weather or climate data.
3. Carry out preliminary analysis to identify potential weather effects.
4. Discuss potential weather effects with statisticians working on road accident statistics and formulate a hypothesis to test.
5. Create appropriate regressors to include in regARIMA modelling.
6. Define a regARIMA model with a full set of regressors to test.
7. Run backwards selection.

8. Analyse model diagnostics and assess stability.
9. Repeat regARIMA modelling as necessary based on results from each iteration.
10. Interpret the results.

Back to flow chart

7.3.2 Select appropriate data

Transport data

What does the dataset contain? The dataset used in this analysis contains administrative records of all police recorded road accidents in Great Britain. The dataset has information on the time and location of the accident as well as information on casualties, the types of road users affected and the severity of their injuries and a number of other variables, including some information on road and weather conditions.

What is the geographical coverage of the dataset? Great Britain.

What is the span of the data? Data are available from January 1979 to December 2012.

Does the data need transforming? To use the regARIMA approach, a time series of counts is required. For this analysis the administrative records were aggregated to produce a monthly time series.

What does the monthly time series measure? The monthly number of killed or seriously injured vulnerable road users in Great Britain. Here, vulnerable road users have been defined as pedestrians, cyclists and motorcyclists.

Met Office data

The weather data selected for this analysis were from the monthly climate series. The monthly climate series are available for the UK but not for Great Britain. One option (that used) is to use the UK data. An alternative would be construct weather data for Great Britain using weather data available for sub-regions of the UK.

What does the series measure? Mean monthly UK temperature.

How frequent is the series? Monthly.

How long is the series? January 1910 to June 2014.

What is the geographical coverage of the series? UK

What other weather data could be used? If analysis was being carried out on individual accidents, then daily or hourly data from the Met Office could have been used.

Back to flow chart

7.4 Assess the data quality and structure

7.4.1 Road accident data

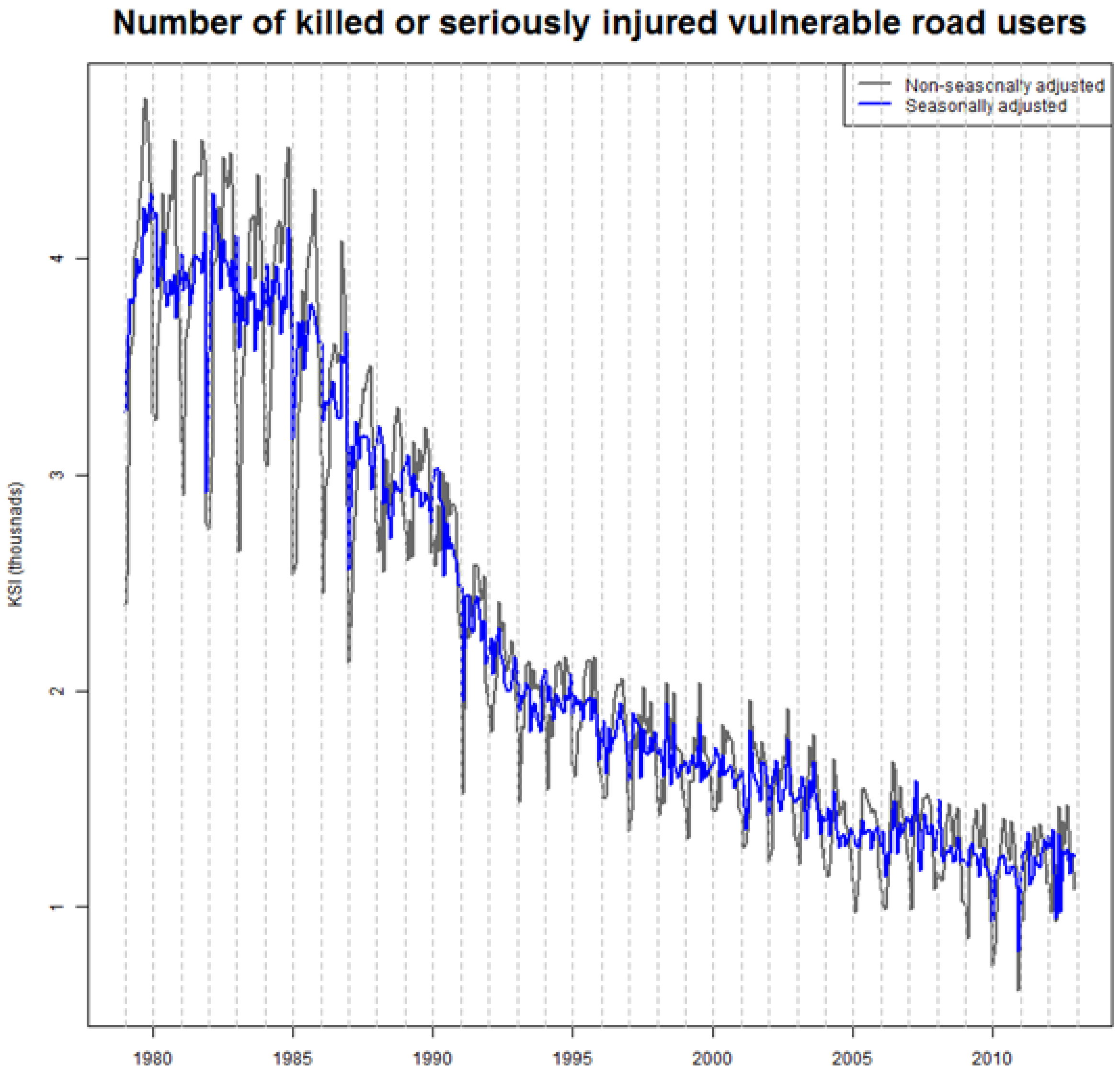
How is the series produced? The STAT19 dataset provides records of all police recorded accidents since January 1979. More information on this dataset can be found here. For this analysis, the casualty dataset has been aggregated to provide monthly counts of numbers of killed or seriously injured vulnerable road users where vulnerable road users are pedestrians, cyclists and motorcyclists.

What does the series look like? The monthly counts of the number of killed or seriously injured vulnerable road users can be seen in figure 19. Both the original time series and the seasonally adjusted series are plotted. It can be seen that generally the numbers have been decreasing over time. The graph also shows that the data are seasonal, with lower numbers of killed or seriously injured road users in the early months of the year and higher numbers towards the end of each year.

Seasonal adjustment of road accident data. Published road accident series are not seasonally adjusted. As there is a clear seasonal pattern in the monthly series the seasonally adjusted series is presented here.

Back to flow chart

Figure 19: *Time series plot of the number of killed or seriously injured vulnerable road users.*

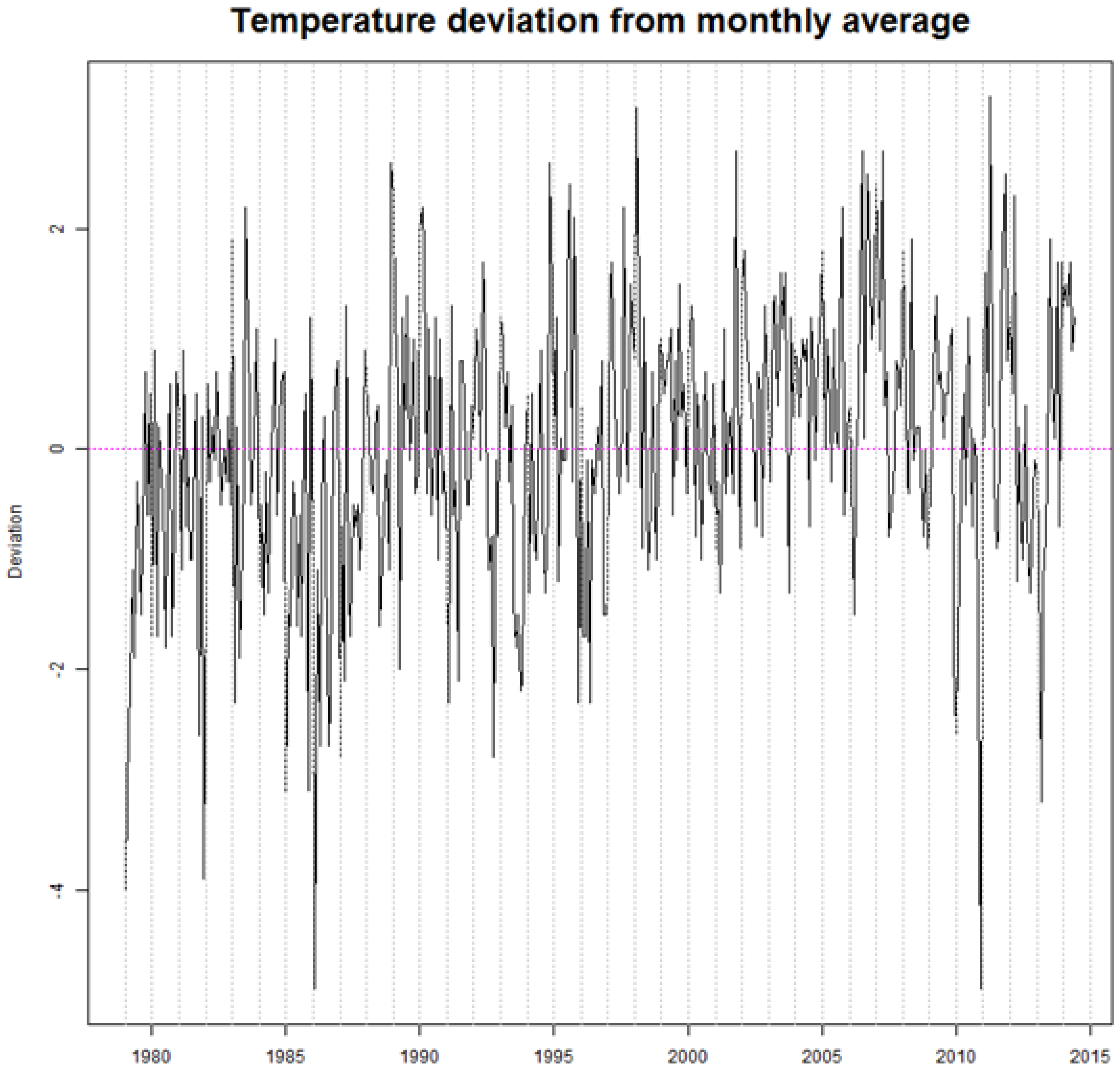


7.4.2 Met Office data

For each month, the average of all of the mean monthly UK temperatures was calculated over the period January 1979 to June 2014. This was then subtracted from the mean UK temperature in that month to produce a time series of temperature deviations from the long-run monthly average. This is plotted in figure 20. The graph shows that colder than average temperatures tend to be more extreme in terms of the temperature deviation, than warmer than average temperatures.

Back to flow chart

Figure 20: *Time series of temperature deviation from long-run monthly average between January 1979 and June 2014.*

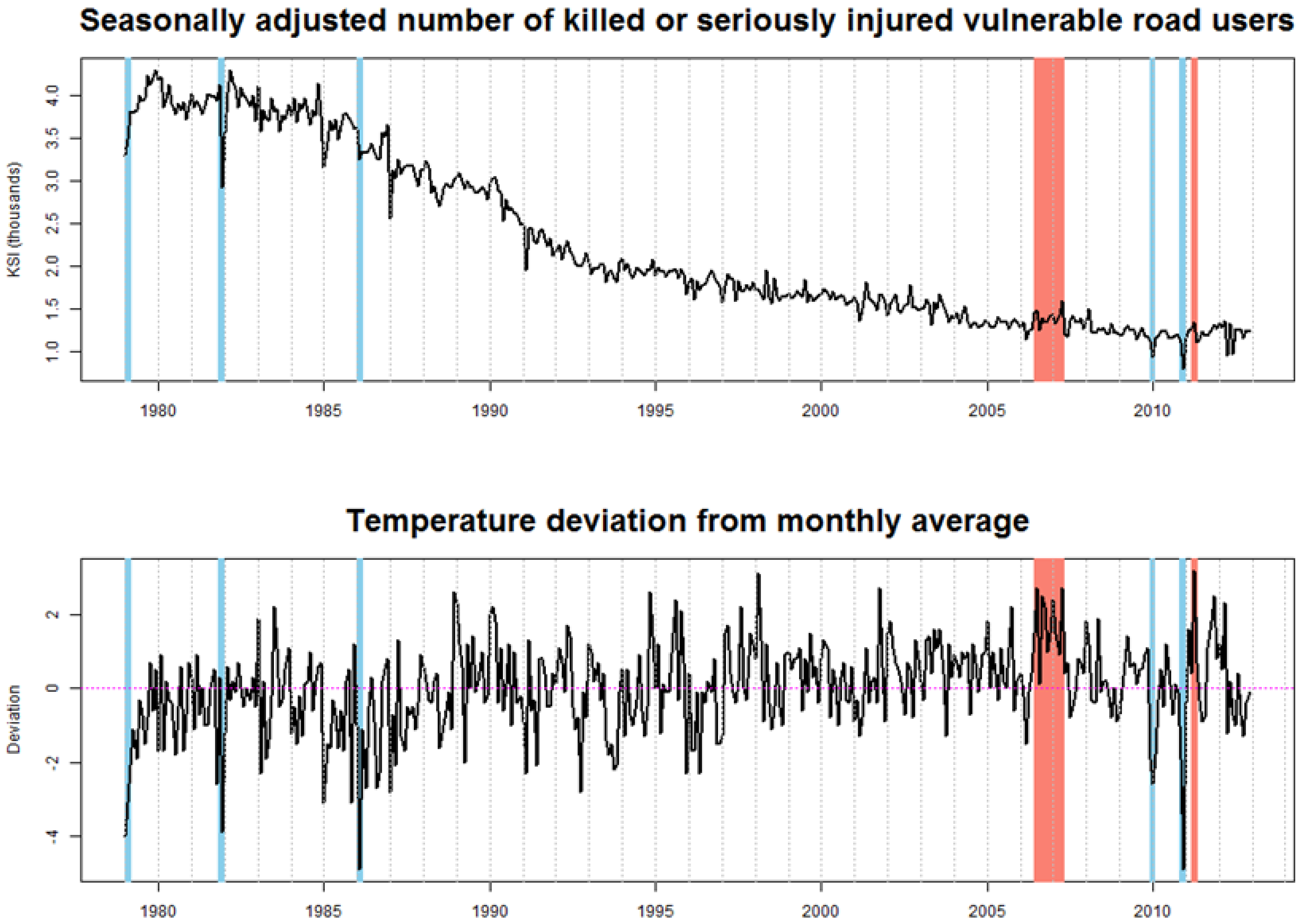


7.5 Initial data analysis

7.5.1 Visualise the road accidents and weather data

In figure 21, both the seasonally adjusted number of killed or seriously injured vulnerable road users and temperature deviations have been plotted. Some blue bars have been used to highlight colder than average months in both plots and red bars for warmer than average. It can be seen that the colder than average temperatures in January 1979 and correspond to large visual dips in the seasonally adjusted killed or seriously injured series. The colder than average weather at the beginning and end of 2010 is also highlighted. This can be seen to correspond with dips in the seasonally adjusted series, but they are smaller than the dips seen in the earlier years of the series. The impact of warmer weather is more difficult to see. The warm March in 2011 corresponds to a visually small peak in the seasonally adjusted series. For most of 2006 and the beginning of 2007, temperatures were higher than average, but it is difficult to see any effect in the seasonally adjusted series.

Figure 21: *Seasonally adjusted killed or seriously injured vulnerable road users and temperature deviation from monthly average.*



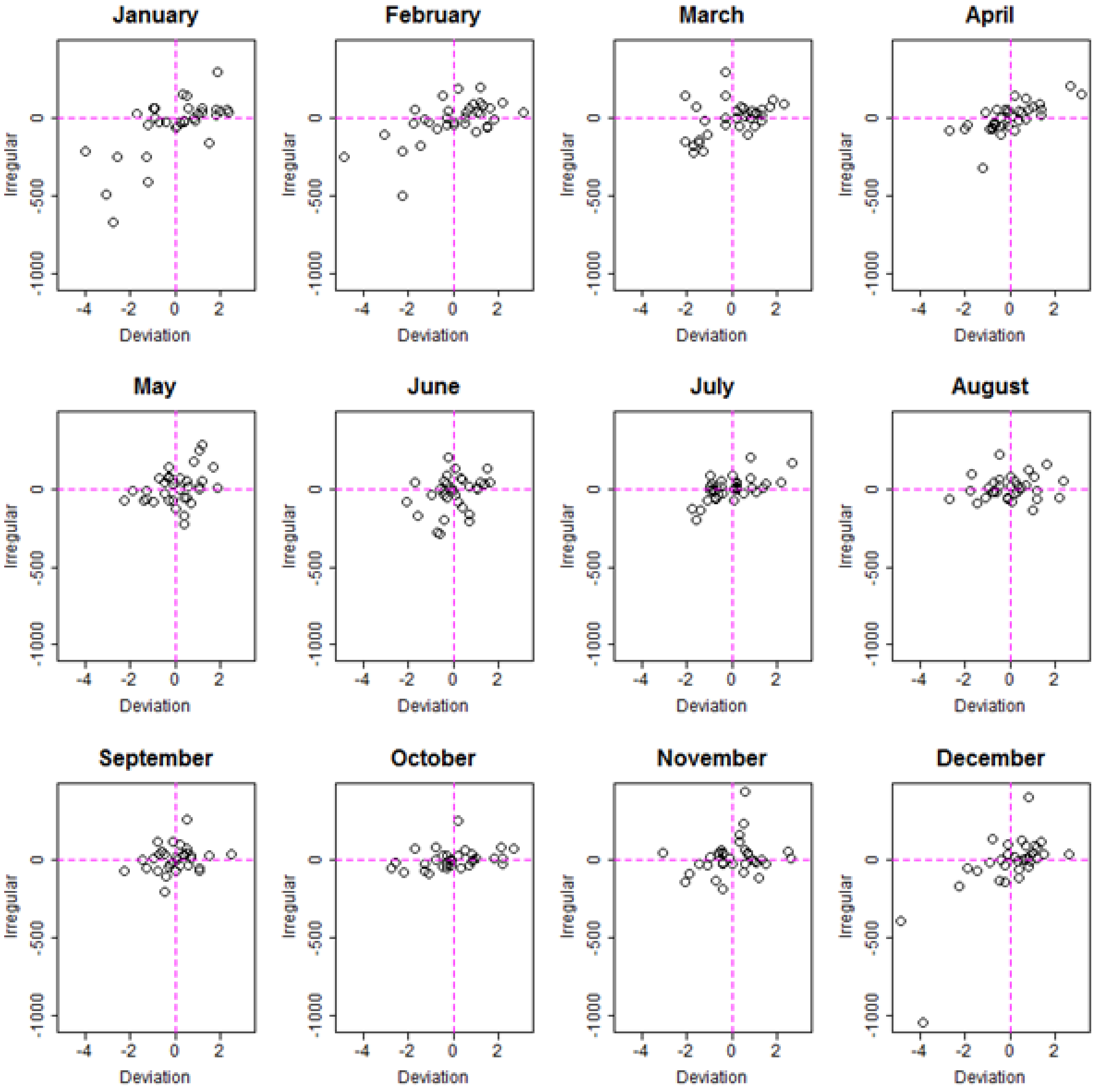
Seasonally adjusted number of killed or seriously injured vulnerable road users

Temperature deviation from monthly average

Figure 22 plots the irregular component of the seasonally adjusted killed and seriously injured vulnerable road users time series for selected months against temperature deviations from the long-run monthly average. There is some evidence of a positive correlation between temperature deviations and unexplained movements in the number of killed and seriously injured vulnerable road users. The effect is more evident in December, January and February where large negative values in the irregular correspond to colder than average temperatures. *Note: an additive decomposition was used to seasonally adjust these series so an irregular value is compared with 0 for no change from the seasonal and trend.*

Back to flow chart

Figure 22: *Irregular component of the killed and seriously injured vulnerable road users time series against temperature deviation from long-run monthly average for selected months.*



7.5.2 Discuss potential weather effects and formulate a hypothesis to test

The visual analysis suggests that there might be a relationship between the number of killed or seriously injured vulnerable road users and temperature. Statisticians at the Department for Transport suggested that warmer and drier conditions might lead to more exposure of vulnerable road users which could lead to more casualties. Given the results of initial analysis and discussions with the Department for Transport, it was decided to use regARIMA modelling to test the effect of temperature on the number of killed or seriously injured vulnerable road users.

Back to flow chart

7.6 Time series modelling

7.6.1 Create appropriate regressors to include in regARIMA modelling

The weather regressors were calculated by taking the deviation in temperature in a given month from the average temperature in that month over the period January 1979 to June 2014. The regressor for a particular month took the deviation if the time period was that month, the negative deviation of the previous month if the time period is the month after, and 0 otherwise. This is illustrated in equation 7.1.

$$x_{i,t} = \begin{cases} tempdev_t & \text{if } month(t) = i \\ -tempdev_{t-1} & \text{if } month(t-1) = i \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Where $tempdev_t$ is the temperature deviation from the long run average at time-point t.

Back to flow chart

7.6.2 Define a regARIMA model with a full set of regressors to test

The time series being analysed is quite long, so the modelling has been carried out on a reduced span of data. In the 1980s the differences between the peaks and troughs in casualties were much larger than in more recent years. Therefore the time series from 1991 to 2012 has been analysed. The number of killed or seriously injured vulnerable road users shows an additive relationship between components of the time series in the most recent years of data, so no transformation of the series is used. The automatic selection of the order of the ARIMA model in X-13ARIMA-SEATS an $(0,1,1)(0,1,1)_{12}$ regARIMA model. Results are presented below.

Back to flow chart

7.6.3 Run backwards selection

The model was run initially with all regressors included. Backwards selection was then used to select a model with a set of regressors which were all statistically significant. Backwards selection is an iterative process in which all of the regressors that are not found to be statistically significant, the least significant is removed. The model is then refitted and the process is run again until all remaining regressors are significant.

The resulting regressors in the final model and their estimated coefficients can be found in the examples of output provided below.

The model has been fitted in Win X-13, R and SAS. The files required to produce these outputs is also provided. *Note: The estimated coefficients vary slightly between the different softwares due to differences in the estimation methods. They may also vary between different versions of the same software.*

regARIMA model output from Win X-13.

```
 -----------------------------------------------
                Parameter  Standard
 Variable        Estimate     Error      t-value
 -----------------------------------------------
 User-defined
 jan              50.9173  15.94284         3.19
 feb              48.7394  13.78862         3.53
 mar              87.1861  16.94539         5.15
 apr              95.2455  16.88851         5.64
 may              77.6860  20.57438         3.78
 jun              81.7293  22.18659         3.68
 jul              54.0150  17.92910         3.01
 oct              27.2009  13.12435         2.07
 dec              61.8352  12.62485         4.90
 -----------------------------------------------
```

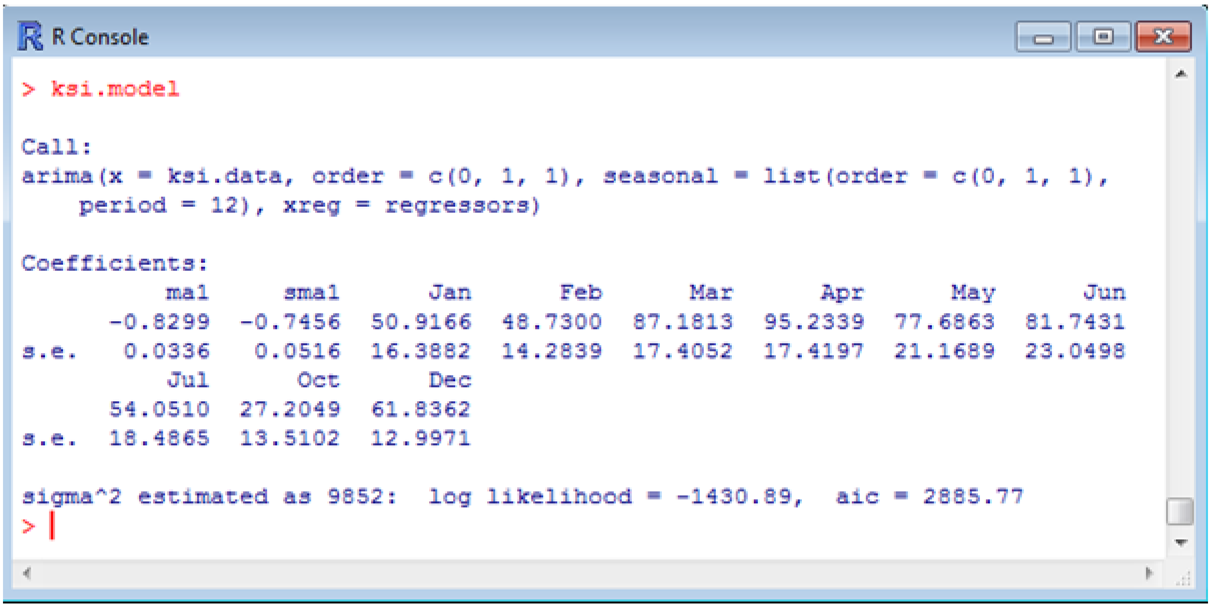
Back to flow chart

7.7 Data and code

Files required to carry out analysis in Win X-13

Right click here to save spec file
Right click here to save data file
Right click here to save regressors file

regARIMA model output from R.

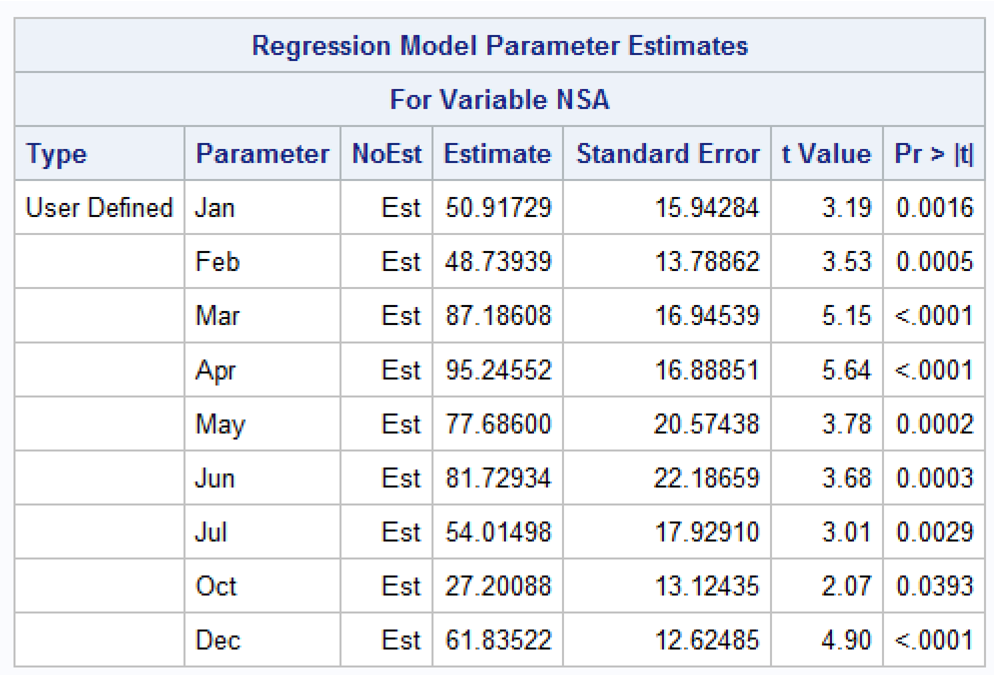


```
> ksi.model

Call:
arima(x = ksi.data, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 12), xreg = regressors)

Coefficients:
          ma1     sma1      Jan      Feb      Mar      Apr      May      Jun
      -0.8299  -0.7456  50.9166  48.7300  87.1813  95.2339  77.6863  81.7431
s.e.   0.0336   0.0516  16.3882  14.2839  17.4052  17.4197  21.1689  23.0498
          Jul      Oct      Dec
      54.0510  27.2049  61.8362
s.e.  18.4865  13.5102  12.9971

sigma^2 estimated as 9852:  log likelihood = -1430.89,  aic = 2885.77
> 
```

Files required to carry out analysis in R

Right click here to save R code file
Right click here to save data file
Right click here to save regressors file

regARIMA model output from SAS.



Regression Model Parameter Estimates

For Variable NSA

Type	Parameter	NoEst	Estimate	Standard Error	t Value	Pr > |t|
User Defined	Jan	Est	50.91729	15.94284	3.19	0.0016
	Feb	Est	48.73939	13.78862	3.53	0.0005
	Mar	Est	87.18608	16.94539	5.15	<.0001
	Apr	Est	95.24552	16.88851	5.64	<.0001
	May	Est	77.68600	20.57438	3.78	0.0002
	Jun	Est	81.72934	22.18659	3.68	0.0003
	Jul	Est	54.01498	17.92910	3.01	0.0029
	Oct	Est	27.20088	13.12435	2.07	0.0393
	Dec	Est	61.83522	12.62485	4.90	<.0001

Files required to carry out analysis in SAS

Right click here to save SAS code file
Right click here to save data file

Back to flow chart

7.7.1 Analyse model diagnostics and assess stability

Likelihood statistics

Win X-13 produces a set of likelihood statistics assessing the fit of the regARIMA model. The estimated likelihood statistics for the road accidents model are provided below. AIC values can be compared for nested models (where all regressors included in one model are included in the other model). Therefore to see if the weather regressors improve the model fit, the AIC of the model including weather regressors can be compared to the AIC of the model with no regressors. The likelihoods of these models are also provided below. The AIC test statistics (highlighted in bold and red) are lowest when the weather regressors are included in the model.

```
Likelihood statistics for model including weather regressors.
-------------------------------------------------------
Number of observations (nobs)                                 264
Effective number of observations (nefobs)                     251
Number of parameters estimated (np)                            12
Log likelihood (L)                                     -1509.0441
AIC                                                     3042.0883
AICC (F-corrected-AIC)                                  3043.3992
Hannan Quinn                                            3059.1131
BIC                                                     3084.3937
-------------------------------------------------------
Likelihood statistics for model not including weather regressors.
-------------------------------------------------------
Number of observations (nobs)                                 264
Effective number of observations (nefobs)                     251
Number of parameters estimated (np)                             3
Log likelihood (L)                                     -1564.2953
AIC                                                     3134.5907
AICC (F-corrected-AIC)                                  3134.6878
Hannan Quinn                                            3138.8469
BIC                                                     3145.1670
-------------------------------------------------------
```

Normality tests

To assess normality of the regARIMA model residuals, three test statistics were calculated; skewness, Gearys a and kurtosis. The results of these tests are provided below. None of these tests indicates a lack of normality in the model residuals. *Note: For a normal distribution, skewness is 0, Geary's a is* $\sqrt{\frac{2}{\pi}}$ *and kurtosis is 3.*

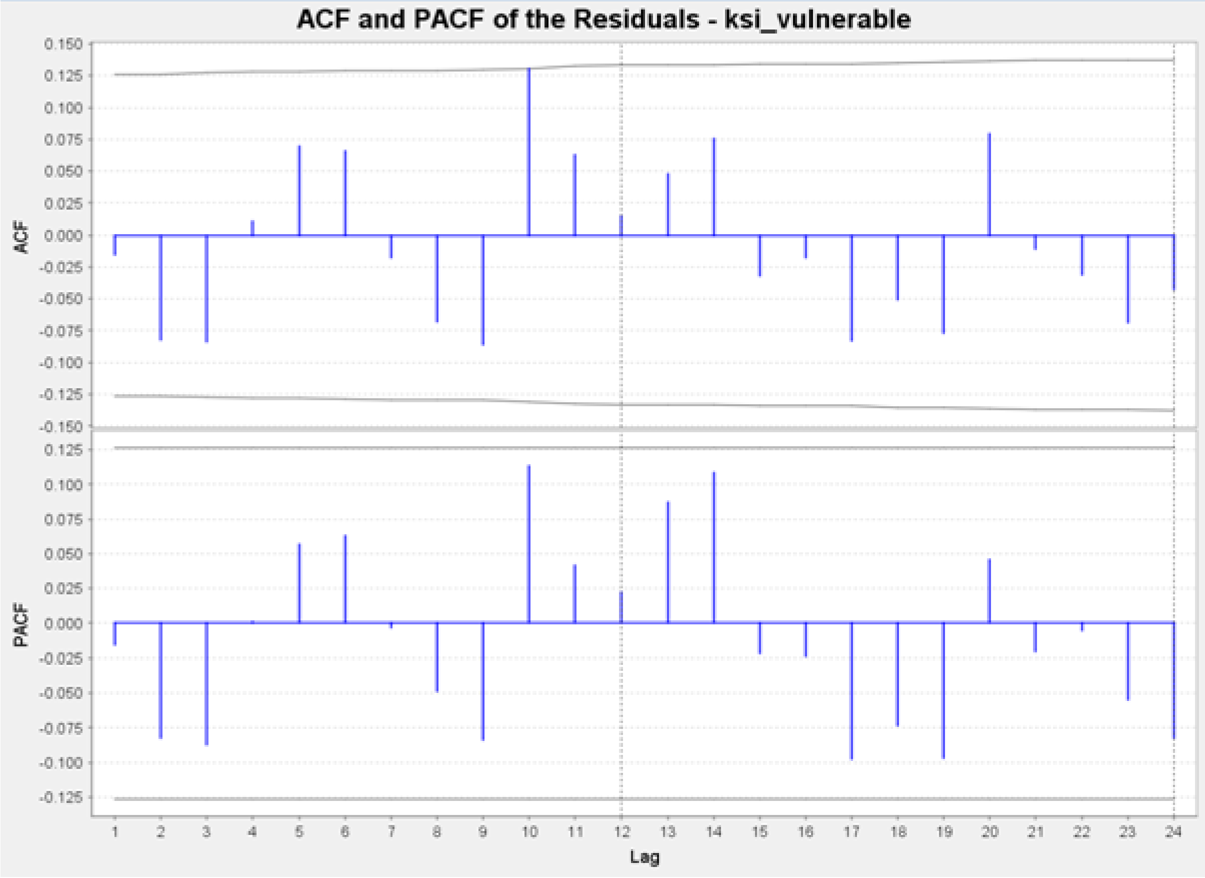
Normality Statistics for regARIMA Model Residuals.

```
Number of residuals    :  251
Skewness coefficient   :  0.1849
Geary's a              :  0.7890
Kurtosis               :  2.9492

No indication of lack of normality
```

Autocorrelation and partial autocorrelation

Figure 23 provide plots of the autocorrelation and partial autocorrelation functions for the fitted model. The plots do not indicate any significant autocorrelation or partial autocorrelation at lags up to 24 months.

Figure 23: *Autocorrelation and partial autocorrelation function of the model residuals.*



The Ljung-Box test statistics on the model residuals have also been provided. As explained in the output, and provided below, for lags with positive degrees of freedom, p-values of less than 0.05 indicate model inadequacy. The values below do not indicate any model inadequacy.

Sample Autocorrelations of the Residuals with the Ljung-Box diagnostic.

```
Lag      1      2      3      4      5      6      7      8      9     10     11     12
ACF  -0.02  -0.08  -0.08   0.01   0.07   0.07  -0.02  -0.07  -0.09   0.13   0.06   0.01
SE    0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.07   0.07   0.07
Q     0.06   1.79   3.59   3.62   4.88   5.99   6.07   7.27   9.22   13.7  14.75   14.8
DF       0      0      1      2      3      4      5      6      7      8      9     10
P        0      0  0.058  0.164  0.181    0.2  0.299  0.297  0.237   0.09  0.098  0.139

Lag     13     14     15     16     17     18     19     20     21     22     23     24
ACF   0.05   0.08  -0.03  -0.02  -0.08  -0.05  -0.08   0.08  -0.01  -0.03  -0.07  -0.04
SE    0.07   0.07   0.07   0.07   0.07   0.07   0.07   0.07   0.07   0.07   0.07   0.07
Q    15.41  16.95  17.22  17.31  19.18  19.87   21.5  23.22  23.26  23.52  24.83  25.33
DF      11     12     13     14     15     16     17     18     19     20     21     22
P    0.164  0.152  0.189   0.24  0.206  0.226  0.205  0.182  0.226  0.264  0.255  0.282
The P-values approximate the probability of observing a Q-value at least
this large when the model fitted is correct.  When DF is positive, small
values of P, customarily those below 0.05, indicate model inadequacy.
```

Seasonality

Win X-13 includes the QS statistic which tests the hypothesis of no seasonality. This test is carried out for the regARIMA model residuals and the results are provided below. These tests show that there is no seasonality in the residuals.

```
QS Statistic for regARIMA Model Residuals (full series):                  0.18
                                                             (P-Value = 0.9142)
QS Statistic for regARIMA Model Residuals (starting 2005.Jul):            0.00
                                                             (P-Value = 1.0000)
```

Back to flow chart

7.7.2 Repeat steps 5 to 8 as required

Having carried out the analysis, further ideas for analysing this data are as follows.

- Use separate regressors for warmer and colder deviations. This could be carried out in a similar way to the example here by defining two regressors for each month, one which takes the value of that months regressor in the previous example if it is greater than 0, and 0 otherwise, the other taking the value of that months regressor if it is less than 0, and 0 otherwise. This would result in 24 regressors being included in the original model.
- Create a variable combining temperature and rainfall to test effects of combinations such as warmer and drier weather.

- Explore interactions between the weather and holiday effects (for example Easter and Bank holidays) or trading days.

- Investigate threshold variables, for example does an effect start only when temperatures reach a certain point?

- Use climate summaries form the Met Office to identify months with snow. Include these months as outliers in the regARIMA model. This can test for whether there is a significant difference in the series in that month.

- Analyse daily or hourly accidents with daily or hourly weather information. This would require a different approach from regARIMA modelling. Some modelling approaches more appropriate to this set up are provided in the literature review section.

Back to flow chart

7.8 Interpretation and communication of the results

The final model includes regressors for the months of January, February, March, April, May, June, July, October and December. The model suggests that as temperatures increase, the number of killed or seriously injured vulnerable road users increase. The effect varies between months, from a 1°C increase in temperature above average (all else being equal) resulting in an additional 27 killed or seriously injured vulnerable road users in October to an additional 95 in April. Similarly, according to the model a 1°C decrease in temperature below average results in between 27 fewer killed or seriously injured vulnerable road users in October and 95 fewer in April. The effect has been modelled as linear so that, for example in October each 1°C increase in temperature above average will see an additional 95 killed or seriously injured vulnerable road users, so if an April is 2°C warmer than the average temperature for April, then all else being equal, there will be an additional 190 more killed or seriously injured vulnerable road users than if the temperature had been average.

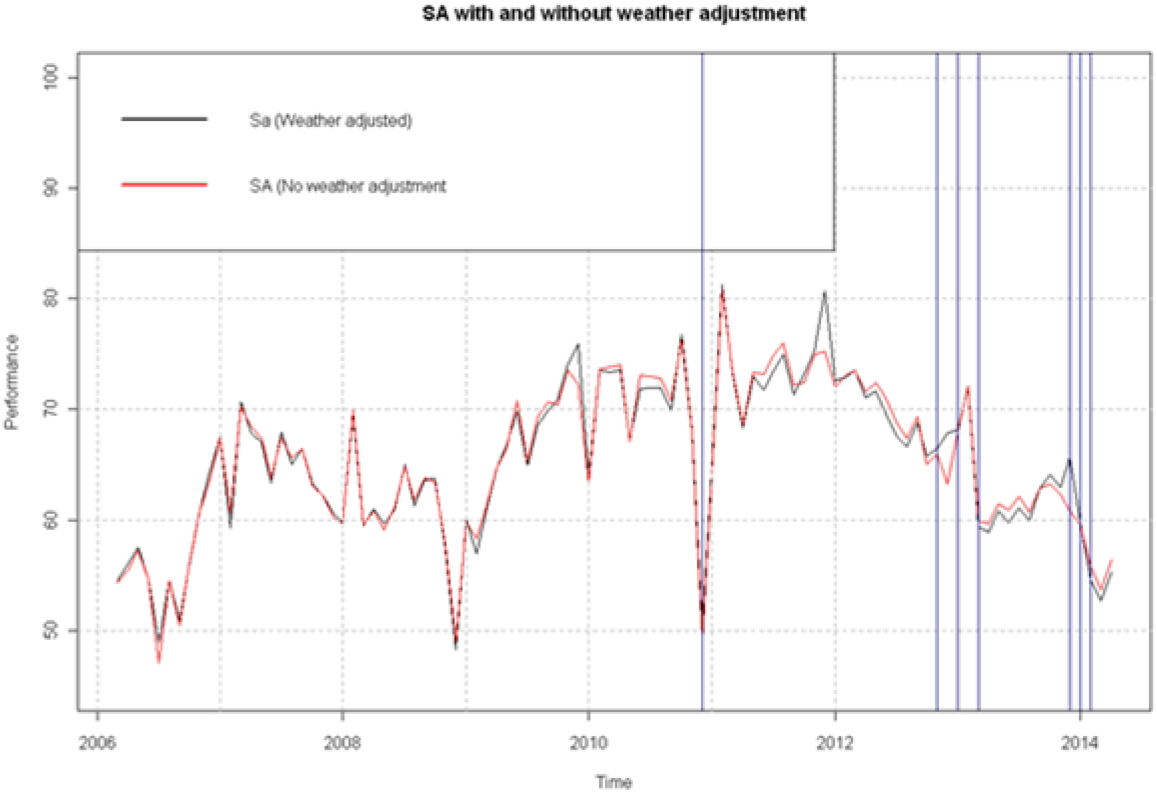
Back to flow chart

8 Ambulance response times

8.1 Overview

Analysis of ambulance response time data looked at weather and climate effects on ambulance performance in Cardiff. The ambulance response time series was compiled from ambulance performance data provided by the Welsh Government. The ambulance data provided was the monthly number of category A (immediately life threatening) calls and the number of category A calls which arrived at the scene within 8 minutes. The data provided by the Met Office included monthly data on temperature, hours of sunshine and days of air frost from the Bute Park weather station in Cardiff. Logistic regARIMA modelling was used to test for weather effects on the proportion of calls where an ambulance arrived within 8 minutes. No significant weather effects were found on ambulance performance when looking at the data on a monthly basis. Further work could be done to extend the analysis to daily data as aggregating weather effects over a month may mask potential effects. It should also be noted that the effects may differ if a rural area were to be considered or if additional variables such as snow days were used.

Figure 24: *Seasonally Adjusted ambulance performance in Cardiff with and without weather adjustments.*



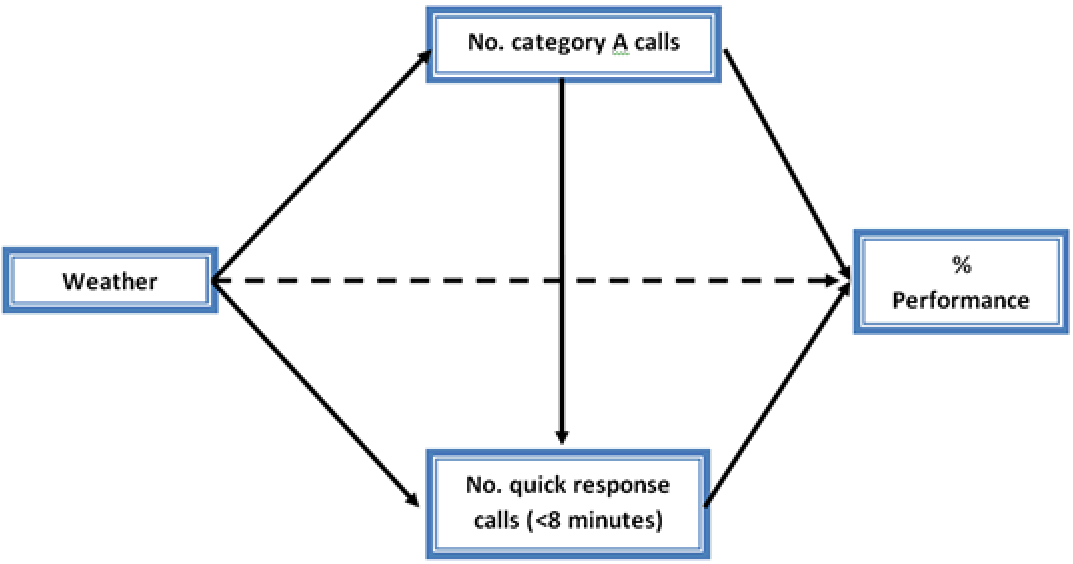
Back to flow chart

8.2 Introduction

The data available for this analysis were ambulance performance data provided by the Welsh Government and weather data provided by the Met Office. The ambulance data provided were the monthly numbers of category A (immediately life threatening) calls and the number of category A calls which arrived at the scene within 8 minutes. The data provided by the Met Office included monthly data on the mean maximum daily temperature, mean daily minimum temperature, hours of sunshine and days of air frost.

One problem with this analysis is that the variable of interest is a percentage. Thus the weather could be thought to affect the numerator, denominator or both. See figure 25.

Figure 25: *Directed graph for ambulance analysis.*



Other problems that could affect weather effect analysis include the following.

- Weather geography - It is important to consider to which geographic area the weather data relates to. For example, using the UK weather observations would not be very useful if you were analysing injuries on Snowdon where the weather is far more variable and extreme than the UK average;
- Series geography - The geography of the series should be considered. For example, the retail sale of petrol is distributed across the UK and so the UK average weather may be used in the analysis. However, if analysing retail sales of designer clothes, most of these sales would presumably be in the city centres of London, Manchester, Birmingham, and so on and so perhaps it would be better to use the weather observations for these areas rather than the UK average;
- Weather Data frequency - The weather data may be available as monthly or daily averages. This will confound analysis of high frequency effects such as squalls (short sharp storms), short heavy showers or long periods of high/low temperature etc;

- Statistics frequency - The statistics will usually be available as monthly or quarterly totals or averages but this will confound analysis of high frequency effects such as a sudden drop in accidents within a day as people react to a squall by delaying their journey;

- Non-linear effects - The effects of weather are not likely to be linear. For example the accident rate may increase with rain up to a point, after which increasing rain is associated with decreasing accidents as people drive more carefully, the accident rate may decrease faster after another point as people delay travel in response to extreme rain;

- Interactions - The effect of weather is likely to involve interactions. For example the sale of barbecue supplies may not be affected by temperature or precipitation alone, but it is likely to be effected by their interaction (warm and dry weather).

- Autocovariances - The effect of weather is likely to be dependent on previous weather. For example, the sale of clothing may decrease if there is rain as people may not want to go out shopping. However if there has been rain for several days people may be forced to go out shopping as they have delayed their purchases for as long as possible.

- Exposure and relative risk - People will probably change their behaviour in response to the weather and this may change the exposure and/or relative risk of an event. For example, with increasing rain people may delay their journeys, reducing the exposure to accidents. Also, the people who do not change their behaviour in response to the weather may be less safety conscious and so the relative risk of accidents may increase.

- Seasonality - The medium to long term variation in the weather is generally predictable and of an annual period (seasonal). This variation will be difficult to separate from other routine annual variation such as financial/tax year start and end, and school holiday timing. This seasonality will be assessed by seasonal adjustment. This means that the analysis of the weather effects will be restricted to extreme weather or deviations from expected weather.

- Expectations - People may respond differently to the weather depending on the time of year. For example hot weather in July may increase ice cream sales but hot weather in December may not.

- Forecast effect - People may react to the weather forecast rather than the actual weather. For example the purchase of barbecue equipment following a forecast of hot dry weather even though the observed weather was cold and wet.

Taking account of all these issues would be complex but the mathematical modelling process proceeds in a piecemeal fashion beginning with a simple model and gradually increasing the complexity of the model. The modelling process to date is summarised below.

- Model 1.0 analysed linear non-interacting weather effects as deviations from the monthly long-run average. This model found that the best model had only taken account of the number of frost days such that the effect of each additional frost day (over the long-run average) was associated with a reduction in ambulance performance of 0.56 percentage points. This modelling process investigated the following.

- Whether the response variable (ambulance performance) can be modelled as a binomial distribution and if so Generalised Linear Model (GLM) regression with a logit link and binomial family will be used.

- Interactions of the weather effects will be analysed using manual forward stepwise regression in X-13ARIMA-SEATS (X13). X13 has been chosen, as it allows for logistic regARIMA modelling whereas it would not be straightforward to do in R directly.

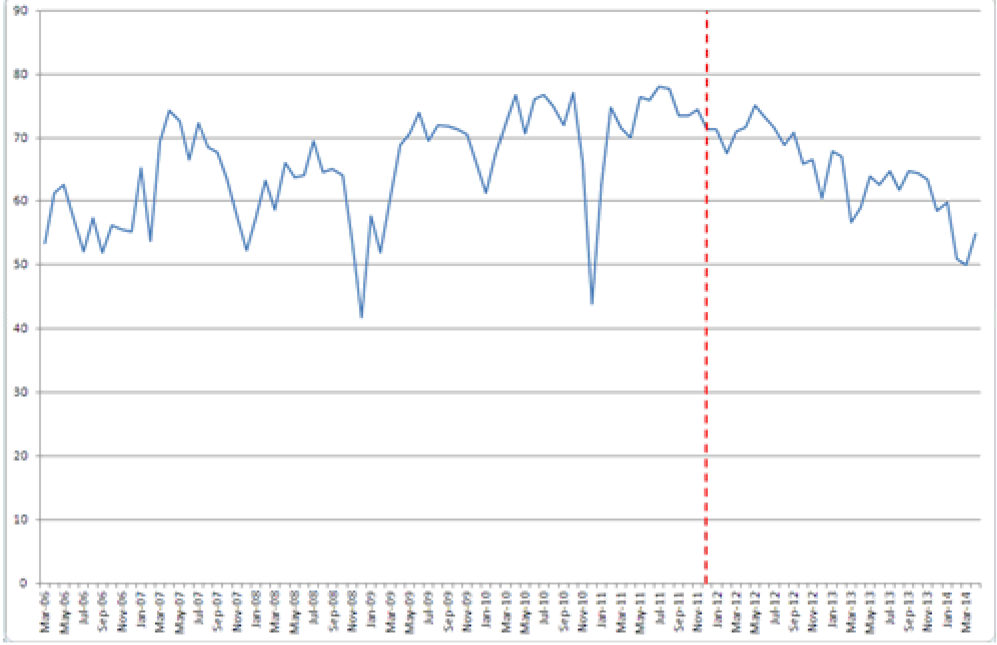
- Regressors will be made for each month, this will allow for weather effects to vary by month. For example a temperature decrease in July is likely to have a smaller effect than a temperature decrease in December which may then be cold enough for icy conditions.

This iteration of the modelling cycle tested the main effects and also the interactions. For interpretation purposes the regression variables will be taken to be deviations from the long-run monthly average, which will allow for interpretations such as “A wetter than average January is thought to have lead to a drop/rise in ambulance performance.” Also a regressor will be created for each month to allow for the weather effects to vary by month, for example, the effect of minimum temperature is probably minimal in the summer months but will probably be significant in the winter months. The modelling will be done in X13 (through R) as X13 allows for logistic regARIMA modelling through the ’transform=logistic’ spec. X13 has software limits on the number of regression variables and so the stepwise modelling will have to be done manually. The way the stepwise modelling will be done will be to test the main effects one at a time (because of software constraints) to find the significant main effects. Then the admissible two-way interactions will be tested. An interaction is admissible if the lower order interactions and main effects to which it relates are also in the model. If there is a constraint on the number of variables due to the software limits, the variables will either be dropped according to the least significant or the variables may be grouped according to similar coefficient values, for example, if the coefficient for the effect of rain is similar for the winter months, the three winter rain regressors could be replaced with one regressor.

The ambulance performance data provided is shown in figure 26 below. The dashed red vertical line highlights a discontinuity in the data due to a change in definitions.

The ambulance data cover the Welsh Local Health Boards (LHB) (see below). The smallest LHB is Cardiff and Vale University and the data have been disaggregated further

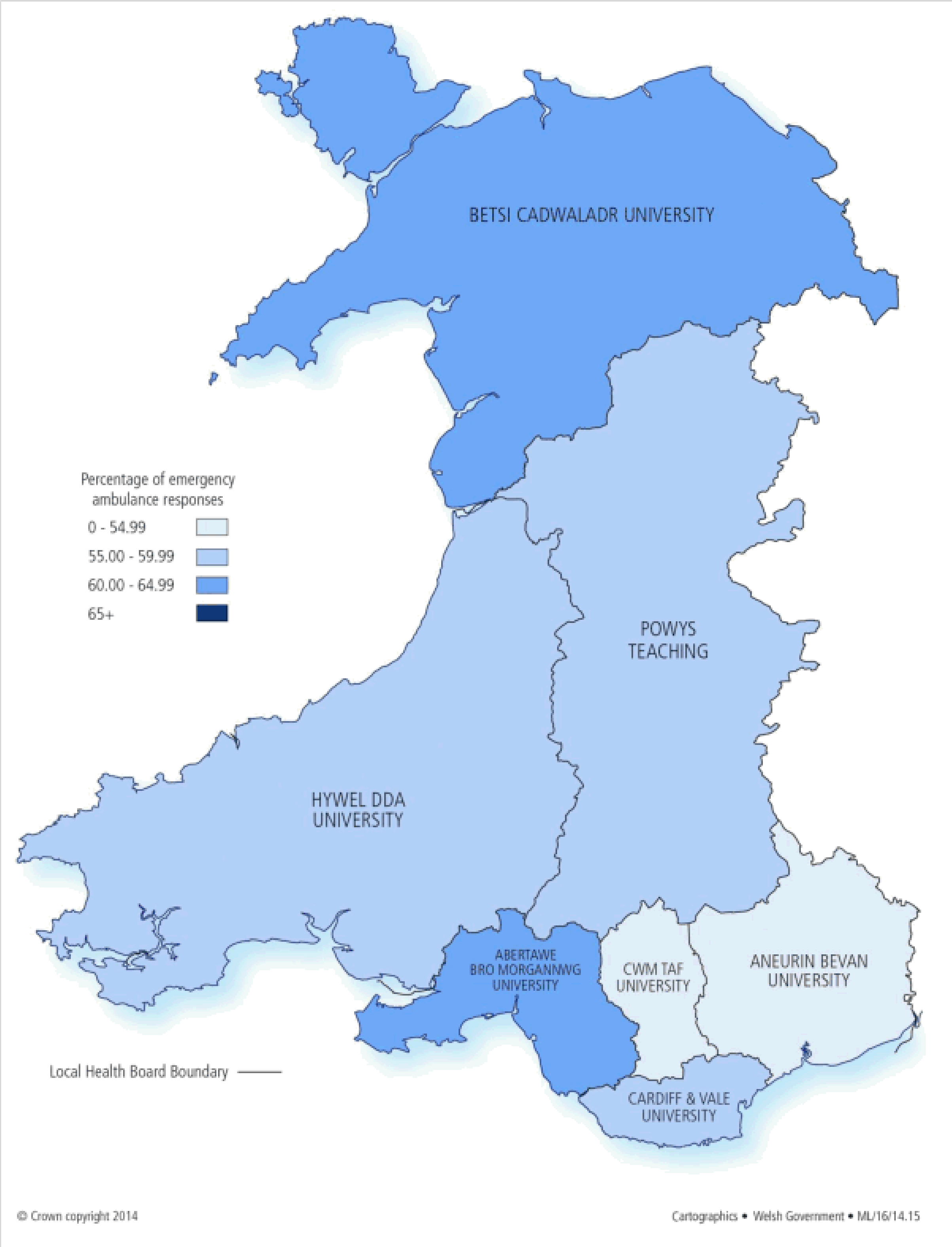
Figure 26: *Ambulance performance data.*



to Cardiff and so this series will be used in the analysis as the weather data relate to the Bute Park observation station (see below).

Back to flow chart

Figure 27: *Welsh Local Health Boards.*


BETSI CADWALADR UNIVERSITY
POWYS TEACHING
HYWEL DDA UNIVERSITY
ABERTAWE BRO MORGANNWG UNIVERSITY
CWM TAF UNIVERSITY
ANEURIN BEVAN UNIVERSITY
CARDIFF & VALE UNIVERSITY
Percentage of emergency ambulance responses
0 - 54.99
55.00 - 59.99
60.00 - 64.99
65+
Local Health Board Boundary

© Crown copyright 2014

Cartographics • Welsh Government • ML/16/14.15

Figure 28: *Welsh Local Authorities.*

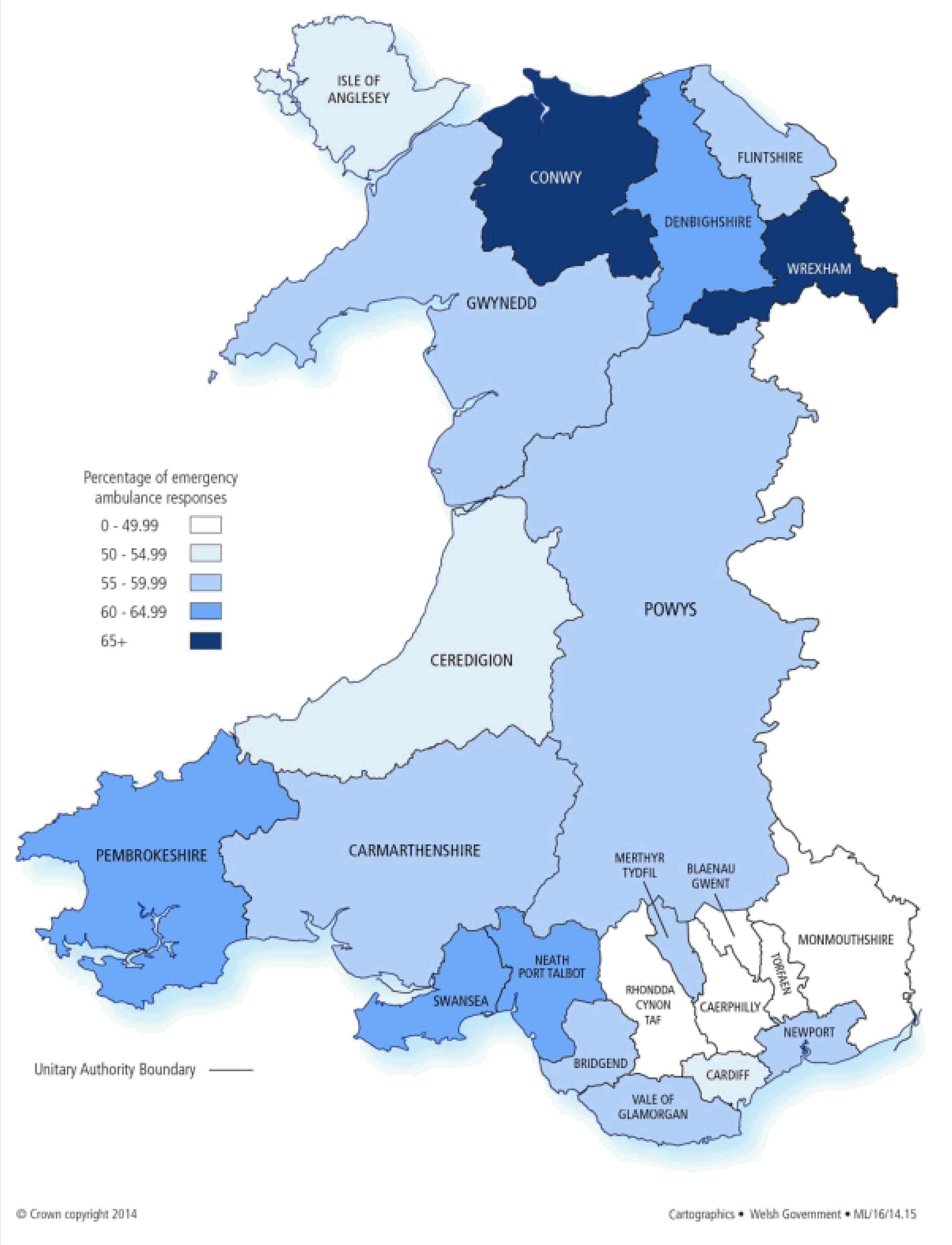
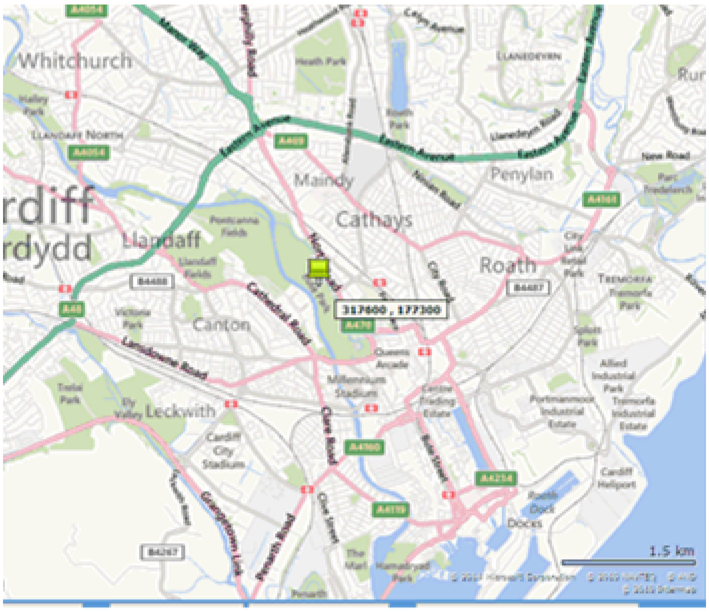


Figure 29: *Observation Station, Bute Park, Cardiff.*



8.3 Analysis

The previous model considered linear non-interacting effects which were deviations from the long-run average. This second iteration of the mathematical modelling cycle will build on the recommendations of the first iteration by:

- investigating whether the ambulance performance (percentage of category A calls responded to within 8 minutes) follows a binomial distribution. If this is so, then the model will use GLM regression with a logit link and binomial family;
- allowing the weather effects to vary by month by creating a regressor for each weather variable for each month. This regressor will be the variables' deviation from the long-run average for that particular month; and
- allowing the weather variables to interact.

The response variable, as before, will be the percentage of category A calls where the 8 minute target was met for Cardiff.

As the setting of this analysis is Official Statistics the model used for this first stage will be a seasonal adjustment model with the weather effect as priors. The coefficients and significance of the regressors will be estimated within X13, as it allows for logisitic regARIMA modelling directly.

The assumptions in this model are:

- the ambulance series has identifiable seasonality;
- the weather effects are linear;
- the weather effects are not confounded by the frequency of the ambulance or weather data;
- the weather effects each month independently, that is to say there are no co-variances between the weather effects;
- the effect of the weather does not effect the exposure (more ambulance calls) or relative risk (more/fewer calls in rush hour traffic);
- the weather forecast does not affect the ambulance series;
- the discontinuity in December 2011 in the ambulance series can be ignored.

The response variable is the percentage of category A ambulance calls where the 8 minute arrival target was met for each month for Cardiff.

The independent variables are:

- mean daily maximum temperature (C);
- mean daily minimum temperature (C);
- days of air frost; and
- total rainfall (mm).

Note that the total hours of sunshine variable is not used as there are no data from April 1996 onwards.

Also, the variables will be the deviation from the long-run average for each month. Using deviation from the long-run mean will not affect the model much as subtracting the mean will simply affect the constant and not the regression variable (see equation 15).

$$\begin{gathered} y = \beta x + \alpha \\ y = \beta(x - \bar{x}) + \alpha = \beta x + (\alpha - \beta \bar{x}) \end{gathered} \tag{15}$$

Although using deviations will not affect the model much, it will affect the interpretation. The interpretation using a model without deviations would be, for example how each millimetre of rain affects ambulance performance; however, the interpretation needed in this modelling is how abnormal weather affects the ambulance performance, for example, a wetter than average July has caused a X% drop in ambulance performance. The model assumption is that the weather affect can be modelled as a prior in a seasonal regARIMA model. Therefore the mathematical relationship between seasonally adjusted ambulance performance and weather effects is as in equation 16.

$$\begin{aligned} \phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D(y_t - \sum_i \beta_i x_{it}) &= \theta_q(B)\Theta_Q(B^s)a_t \\ y_t = &\text{ Observed series} \\ x_{it} = &\text{ Regressor (location/size of weather effect)} \\ \beta_i = &\text{ Coefficient of regressor (size and effect)} \\ Bz_t = z_{t-1} &\text{ The backshift operator} \\ s = &\text{ Seasonal period} \\ \phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p & \\ \Phi_P(B^s) = 1 - \Phi_1 B^s - \ldots \Phi_P B^{Ps} & \\ \theta_q(B) = 1 - \theta_1 B - \ldots - \theta_p B^q & \\ \Theta_Q(B^s) = 1 - \Theta_1 B^s - \ldots \Theta_P B^{Qs} & \\ d = &\text{ Order of nonseasonal difference} \\ D = &\text{ Order of seasonal difference} \\ a_t \sim iid(0, \sigma^2) & \end{aligned} \tag{16}$$

In this model, if analysis of the response variable (ambulance performance) justifies it, a logistic regression model will be used. Therefore the assumed relationship between non-seasonally adjusted ambulance performance and the weather effects is as in equation 17.

$$\begin{aligned} \ln(\frac{y_t}{1-y_t}) &= \sum_i \beta_i x_{it} \\ y_t = &\text{ Percentage of category calls responded to within 8 minutes} \\ \beta_i = &\text{ Weather effect for variable } i \\ x_i = &\text{ Regressor for weather effect } i \end{aligned} \tag{17}$$

Note that as there will be one regressor for each month, and there are four main effects there will be 48 main effect variables. This is a lot of variables when there are only 98 observed data points but as logistic regression is being used the total number of observed 'trails' (ambulance calls) varies from 997 per month to 1842 per month.

For this model the weather regressors will simply be deviations from the monthly average for each month. The monthly averages will be calculated using the full series of weather data to provide the long-run average. They are calculated as in equation 18.

$$\begin{aligned} x_{rain,June2009} &= O_{rain,June2009} - \frac{\sum_{t=June2006}^{June2013} O_{rain,t}}{N_{June}} \\ x_{rain,June2009} = &\text{ Regressor for rain in June 2009} \\ O_{rain,June2009} = &\text{ Observed rain (total rainfall (mm)) for June 2009} \\ N_{June} = &\text{ Total number of June observations in time series} \end{aligned} \tag{18}$$

Note that the regressor for June will be zero for every month that is not June and so on. If the regressors were not zero for months which they did not correspond to, the product of the coefficient and the monthly average would remain and this would affect the model and interpretation.

The software used was X-13ARIMA-SEATS version 1.1 build 9. The version of RStudio used in the modelling is version 0.94.92. The version of R used is 2.13.1.

Back to flow chart

8.3.1 Method

- Import data into R and then analyse the percentage of category A calls responded to within 8 minutes to determine the possible distributions.
- Calculate the long-run average for each variable (and interaction) for each month.

- Calculate the regressors as the deviation of the variable from the long-run average.
- Use stepwise forwards logistic regression to determine the best model. This will involve calculating the interactions of the variables as appropriate.
- Test the main effects individually;
- Test the admissible interactions;
- If there is a restriction on the number of variables consider grouping variables or dropping according to the least significant.
- Check the regression assumptions are satisfied for the chosen model.
- Seasonally adjust the series both with and without the weather effects from the chosen model as priors.
- Compare the AIC, forecast error and other diagnostics of the SA series with and without the weather adjustments.
- Qualitatively compare the SA series with and without the weather adjustments paying particular attention to the periods with notable weather events.

Back to flow chart

8.4 Discussion

The response variable was first plotted as a time series in figure 30. The series seems to have seasonal troughs in the winter and peaks in the summer. After the discontinuity in December 2011 the series seems less volatile.

The scatterplot of the variables (figure 31) suggests that the ambulance response performance is most closely related to temperatures with increasing performance with increasing maximum and minimum temperatures. There is multicollinearity in the variables, as would be expected, with the maximum and minimum temperatures and air frost days having high correlation. This suggests that the backward selection will find a model with one of the temperatures or air frost.

The stages of the forward stepwise regression are shown below. Significance is taken to mean a t-value outside (-2, 2).

Figure 30: *Time series plot of ambulance performance*

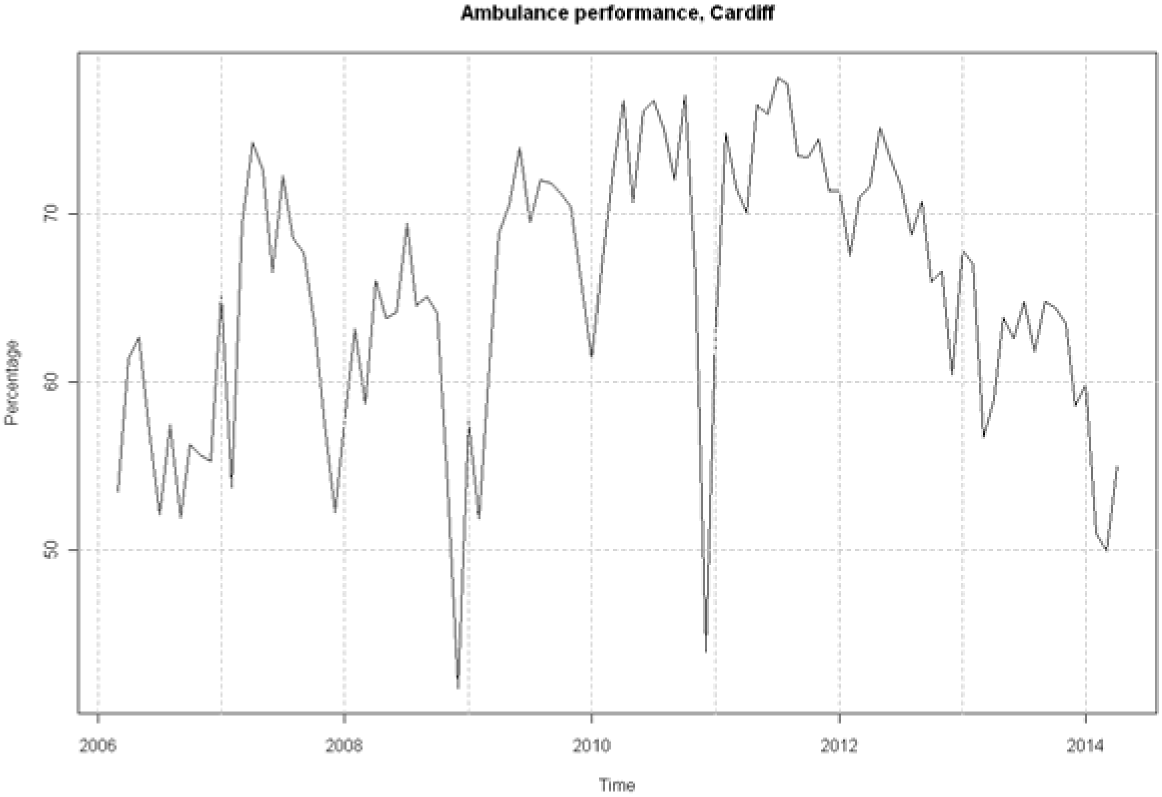
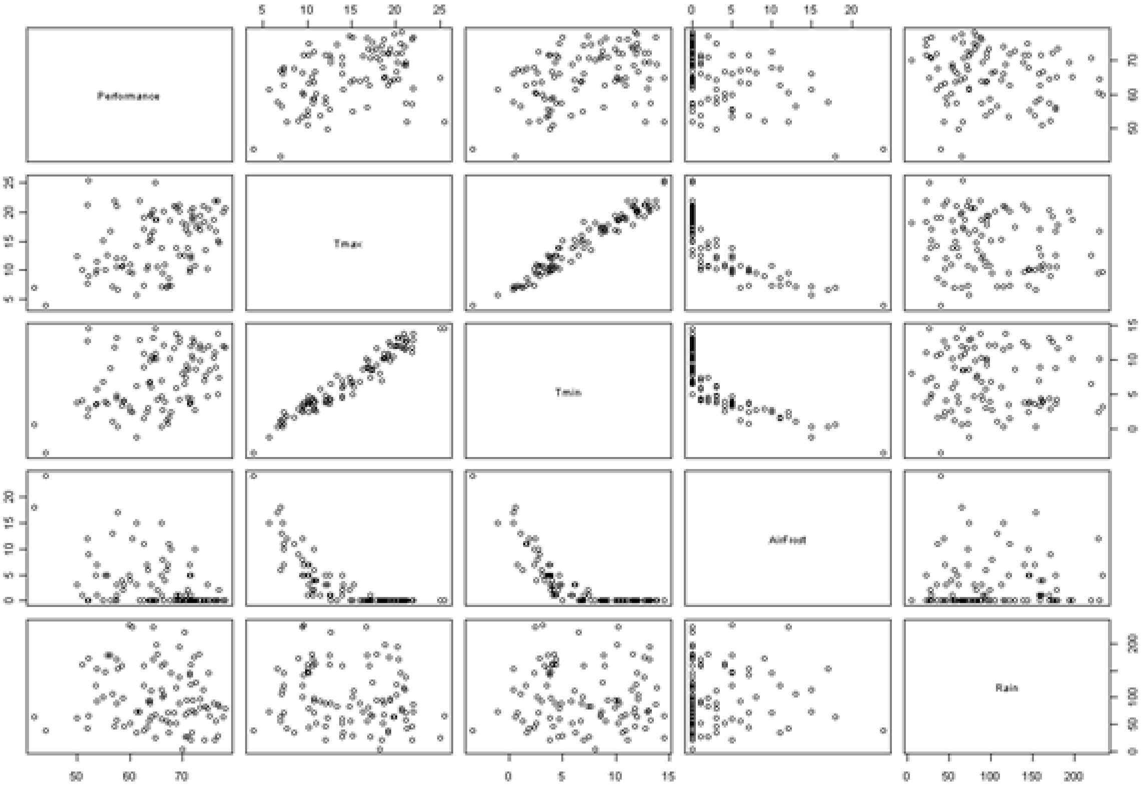


Figure 31: *Scatterplot matrix of variables.*



Step 1: This model tested each month for each main effect. The variables found to be significant were:

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
af	Y	Y	N	N	Y	N	N	N	N	N	N	Y
tmax	Y	N	Y	N	N	N	N	N	N	N	N	Y
tmin	Y	N	N	N	N	N	N	N	N	N	N	Y
rain	Y	N	N	N	N	N	N	N	N	N	N	Y

Y - Variable was found to be significant. N - Variable was found to be non-significant.

Step 2: This model tested the model including all of the above significant main effects. The variables found to be significant were:

	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
af	N	Y	NT	NT	Y	NT	NT	NT	NT	NT	NT	N
tmax	N	NT	N	NT	NT	NT	NT	NT	NT	NT	NT	N
tmin	N	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	N
rain	Y	NT	NT	NT	NT	NT	NT	NT	NT	NT	NT	N

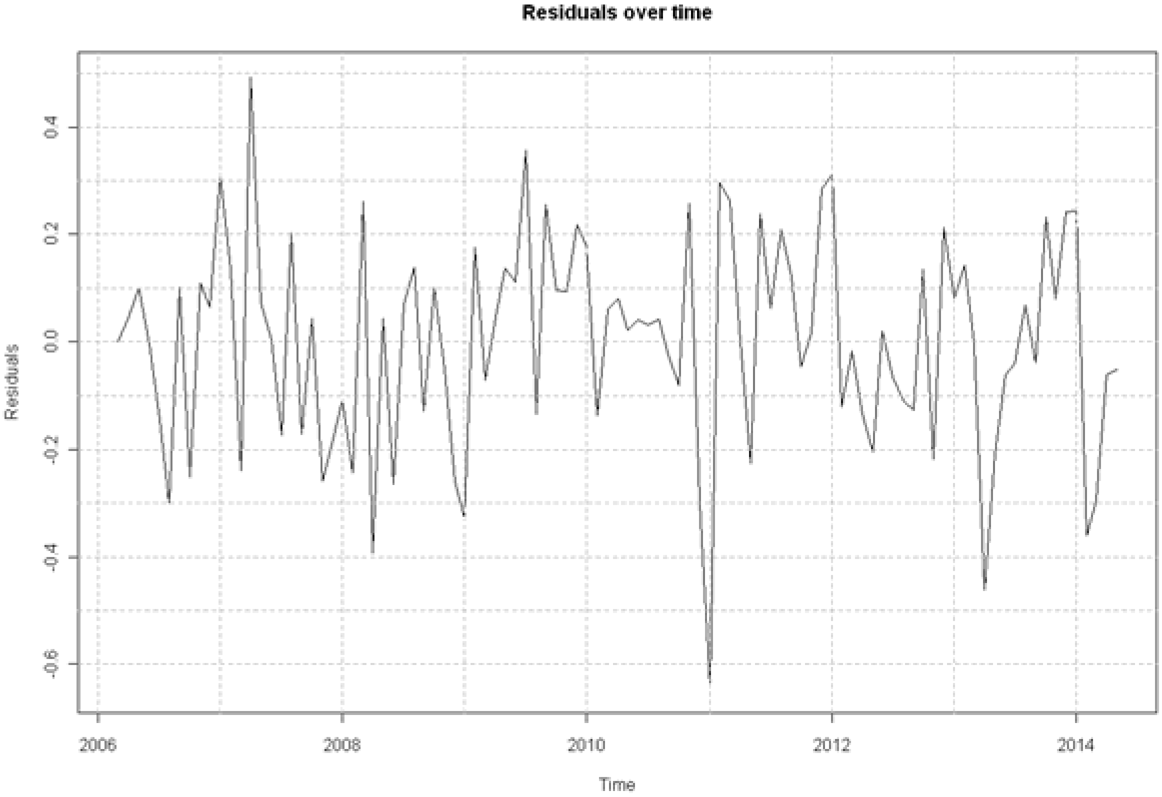
Y - Variable was found to be significant. N - Variable was found to be non-significant. NT - Variable was not tested in this model.

To compare the model results against reality the regression assumptions will be analysed to see if the regression is valid. Also the absolute average within sample forecast error of the seasonal adjustment model with and without the weather prior adjustment will be compared to see if the weather adjustment improves the model fit. A qualitative comparison will also be made of the seasonally adjusted series, both with and without weather adjustments with particular attention paid to periods with notable weather events. The AICC of the models with and without weather effects will be compared to see if the extra complexity is justified. Finally the parameter coefficients, 95% confidence intervals and p-values for each of the weather effects will be presented and interpreted.

Analysis of the regARIMA residuals (figure 32) suggest that there is no evidence of poor model fit. The residuals appear randomly distributed over time and the standardised residuals are normally distributed. The plot of the residuals against predicted values however does suggest that there may be a slight bias to underestimate for higher values.

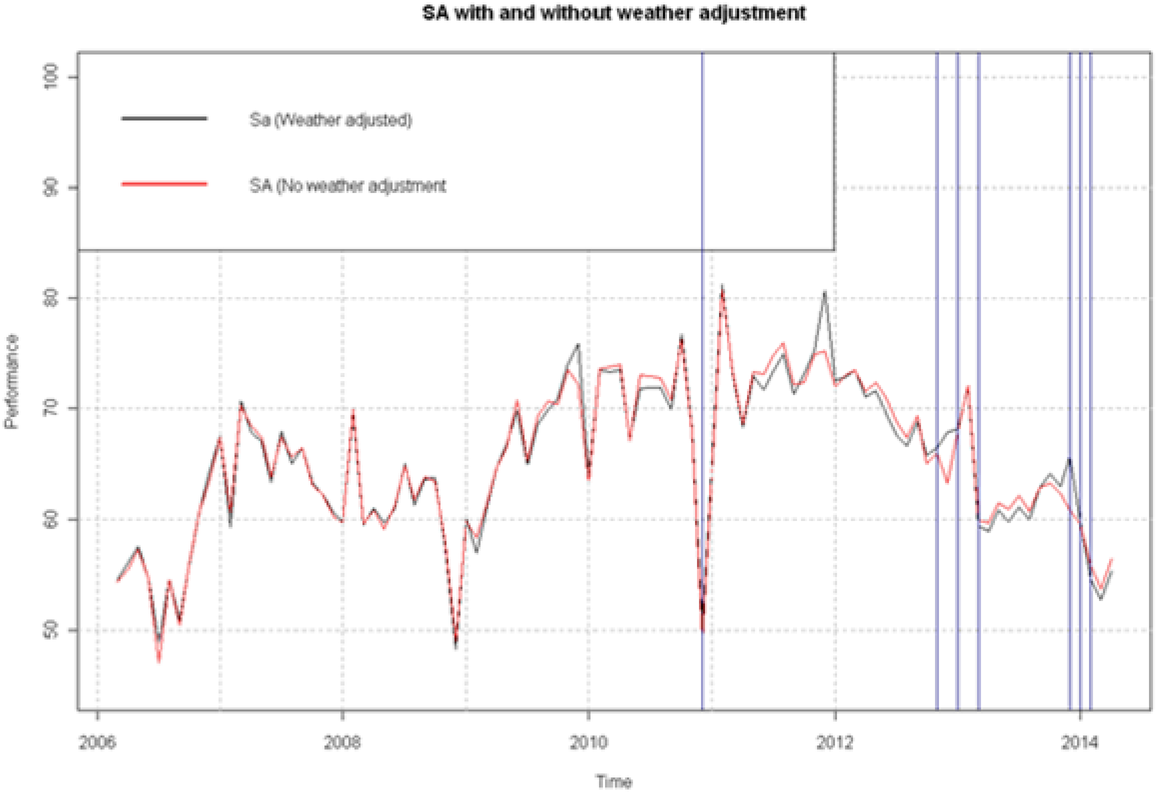
Comparing the SA with and without the weather adjustments suggests that performance was affected by weather in December 2009, 2011 and 2013, with 2013 being a notable weather event (strong winds and heavy rainfall). It is also notable that there was no difference in December 2010 during the extreme snow. This is to be expected, as snowfall was not a weather variable. However, given the steep drop in December 2010 this does

Figure 32: *RegARIMA residuals over time.*



question whether a snow fall variable should be made available. The plot (figure 33) of the seasonally adjusted with and without weather adjustments and also the weather effects included in the model show a curious pattern where the differences between the SA series is not always in a month affected by the weather effects (Jan, Feb or Mar). This should be investigated further.

Figure 33: *SA with and without weather adjustments and weather effects..*



The AICC for the no weather effect model was -296.9119. The AICC for the model with weather effects was -248.781. Thus the AICC prefers the simpler model which does not take account of the weather. The average absolute percentage error of forecasts within the last three years was 6.988 for the no effect model and 7.5965 for the model with the weather effects.

The parameters estimated by the model were:

Parameter	**Coefficient**	**95% CI**	**P-value**
Air frost days. February	-0.0348	(-0.0683, -0.0012)	0.0226
Air frost days. May	0.2648	(0.0117, 0.5179)	0.0216
Rain. January	0.0033	(0.0007, 0.0059)	0.0085

These parameters are to be interpreted as the effects on the log-odds, as logistic transformation is used. So the effect of each air frost day over the long-run average in February is to decrease the odds of an ambulance call meeting the target by 3.4%. The effect of an additional air frost day above the long-run average in May is to increase the odds by 30.3%. Each additional millimetre of rain above the long-run average in January increases the odds of an ambulance call meeting the target by 0.3%.

This iteration of the modelling cycle used logistic regression within X13 to test the main effects and their interactions. The results were that only additional air frost days in February and May and rain in January were significant predictors. However, model comparison with the null model suggested that including the weather effects was not a significant improvement on the null model. Also, the interpretation of the effects was not intuitive, with additional air frost days in May being associated with increased performance. The same was true of rain in January. Also, the largest weather effect from the time series appears to be the extreme snow in December 2010 but there is no weather variable available for this.

Back to flow chart

8.5 Data and code

The code and data used for the analysis in this chapter are provided in the files attached below.
Right click here to save R code file
Right click here to save data file

Back to flow chart

9 FAQs

How long should my time series be? There is no exact answer to this question as it depends on the nature of the data. For example, periodicity, number of time points, stability of the series, discontinuities in the series and so on. If your series is too short you cannot fit many of the models presented in this guide automatically. If your series is too long you may find problems with changing variance or other types of structural changes in the data that are not well dealt with in an ARIMA model. For fitting regARIMA models we have found that about twelve years of monthly data works well.

Can I model annual/quarterly/monthly data? Yes, you can model any periodicity or even mixed frequency data. However, if you are using X-13ARIMA-SEATS you will be limited to using data that have the same periodicity.

Can I have missing data in my time series? Yes. However, depending on your approach to modelling some additional modelling may be required. For example, X-13ARIMA-SEATS does not allow missing data.

Is my model right? No model is right. It is important to have some theory as to why your model specification might be correct. Testing your model with model diagnostics and attempting some sort of out of sample type testing is useful. Models may not be right but they can be useful.

What is an exogenous variable? An exogenous variable is one that is not assumed to be a random variable in the model. For example, in the regARIMA models presented in this guide, the variable we want to model is assumed to be a random variable that follows a stochastic process. The regressors, for example a weather variable is a deterministic variable, and is not assumed to be random. This is described as an exogenous variable.

What is autocorrelation? Autocorrelation is correlation in a time series where time points are correlated with other time points at particular lags. For example, time points t may be correlated with time points t-1. If you found autocorrelation in your model residuals then there is some structure that you can further explain and must be explained in your model to prevent poor inference. ARIMA models have been found to be effective in dealing with autocorrelation.

What is stationarity? Stationarity is where the mean and co-variance structure of a time series does not change with time. A time series is considered to be weakly stationary if the mean and variance are not time varying. For example, if you time series has an upward trend or becomes more or less volatile in different sections of the series then it is non-stationary.

What is a regARIMA model? A regARIMA model includes and ARIMA model and regression part. It can be thought of as and ARIMA model with a time varying mean (the time varying part is due to the changing values of the regressors).

Back to flow chart

10 Acronyms and notation

AIC(C)	Akaike Information Criteria (Corrected)
ARIMA	Auto Regressive Integrated Moving Average
DEFRA	Department for Environment Food and Rural Affairs
DfT	Department for Transport
GSS	Government Statistical Service
ONS	Office for National Statistics
SEATS	Signal Extraction of ARIMA Time Series
TSAB	Time Series Analysis Branch

Back to flow chart

11 Bibliography

CBS. (2014). How unusual weather influences GDP. Retrieved from url

Chatfield, C. (1995). Problem Solving: A statistician's guide second edition. Boca Raton: Chapman & Hall/CRC.

Chatfield, C. (2004). The Analysis of Time Series: An Introduction Sixth Edition. Boca Raton: Chapman & Hall/CRC.

Cox, D. R. (2007). Applied Statistics: A Review. The Annals of Applied Statistics , Vol.1 No.1 1-16.

DECC. (2011, January). Temperature correction of energy statistics. Retrieved December 3, 2014, from url

Eurostat. (2009). ESS Guidelines on Seasonal Adjustment. Retrieved May 12, 2014, from url

Gomez, V., & Maravall, A. (2001). Seasonal adjustment and signal extraction in economic time series. In D. Pena, G. C. Tiao, & R. S. Tsay, A Course in Time Series Analysis. New York: Wiley & Sons.

Harvey, A. (1993). Time Series Models second edition. Harlow: Pearson Education.

Ladiray, D., & Quenneville, B. (2001). Seasonal Adjustment with the X-11 Method. New York: Springer.

Met Office. (2014). Weather and Climate Statistics. Met Office.

ONS. (2014). Guide to Seasonal Adjustment with X-13ARIMA-SEATS.

ONS. (2014). How sensitive to the weather is the retail sector? Retrieved from url

R Core Team. (2013). R: A language and environment for statistical. Retrieved from url

Rahman, S. (2011). Temperature correction of energy statistics. Retrieved from url

Shumway, R. H., & Stoffer, D. S. (2006). Time Series Analysis and Its Applications: With R Examples Second Edition. New York: Springer.

Smith, K. (1982). How seasonal and weather conditions influence road accidents in Glasgow. Scottish Geographical Magazine , 98:2, 103-114.

USCB. (2013). X-13ARIMA-SEATS Reference Manual Version 1.1. Retrieved from url

Back to flow chart

Annex A

Full details on the questions in section 2.1.

What do your data measure? For example, is it a level (such as the number of road accidents in a period) or a rate (such as the proportion of ambulance callouts that meet the target response time in a period)? This may have implications for possible transformations of the data to use in your modelling and also on the way in which you might expect your data to be affected by the weather. For example, indices, such as the Retail Sales Index example in this guide, often (but not always) exhibit a multiplicative behaviour where variation is proportional to the level of the data. When modelling such data, in order to stabilise the variance (in the examples we consider below, a fixed variance is assumed) a log transformation may be appropriate. In the ambulance data, this is a rate and therefore has a strict interval of values between zero and one. An appropriate transformation for such data is the logistic transformation.

Do you have flow or stock data? The examples provided in this guide are all flow data, which is activity over a period (for example, total sales in a month). When starting to think about what weather or climate variables might be appropriate you will need to consider how the weather could affect your series. If you have stock data, it is most important to be aware of the time point at which the stock is measured and also then how the stock may be affected by weather or climate. For example, if you believe there may be a direct relationship between your stock of something measured on the middle day of the month and weather from the previous week then you will need to obtain appropriate weather data for the week in question.

What is the span of your data (start and end date)? For time series modelling explored in this guide it is important that you have a reasonable span of data. This does depend to some extent on how volatile it is, but as a rough guide for the examples below you would want at least five years of monthly or quarterly data (preferably more for the quarterly). If you have a particularly long time series (more than fifteen years of a monthly time series), for the methods used in our examples, you may want to limit the span to the latest fifteen years. We have generally found that about twelve years of monthly data works well for fitting ARIMA models. This can be caused by changes in the autocorrelation structure of the series over longer spans of time. The same may be true for the correlation between weather data and your data. It will be important to test the stability of your model by fitting it to data from different time spans if you have a reasonable length time series.

What is the geographic coverage of your data? This is an important consideration for then selecting appropriate weather or climate data. In general you are likely to want the geographic coverage to be the same, although there may be circumstances where the weather in one region affects a measured outcome in another region. What is the frequency of your data? Most official statistics publications are monthly, quarterly or annual. The models presented in this guide cover monthly and annual examples, but any period could be considered. By default there are some restrictions to the period in some of the software used in these examples, but these could be changed if desired see chapter 7.15 of (USCB, 2013).

What types of weather or climate variables might be expected to affect your data? Before searching through the possible sources of weather and climate data it is important to start building some hypothesis of how weather or climate might affect your data. There may already be some literature on this that will give you an idea of appropriate variables to consider. Section 2.2 provides further information on available sources of data. If the data you require is not described in that guide you can contact the Met Office to enquire about creating bespoke data sets.

How might your data be affected by the weather and climate? Apart from considering the sources of weather data, it is crucial to consider how your data might be affected by weather and climate, for example, could they be affected by rain only, or is there a combination of measured weather or climate variables that might affect your data. This will help in constructing your model. For example, does it make sense to assume a linear relationship between temperature and your time series or would some other relationship be expected. Considering the different types of models you may want to test will help identify the sorts of weather and climate data you will need.

Are there any details of the compilation process of your data that need to be considered in your analysis and in your selection of appropriate weather variables? Thinking about the production process of your data is important as it may have implications for the structure of the data (from a time series perspective) and the relationship to weather data. For example, in the section on retail sales we briefly discuss the fact that retail sales data are not collected on a strict calendar month basis, although they are published on a calendar month basis. Either the data should be calendarised or the weather data should match the retail sales period in the data.

How are your data currently published, and how might your publications be affected by the analysis? Conducting an analysis may be of interest only to the producers of the data, to help them understand what sort of things might affect their series. However, users of the data may also find the analysis of use. One of the recommendations in the Eurostat Guidelines on Seasonal Adjustment (Eurostat, 2009) is not to adjust a time series for weather effects. Clearly for some seasonally adjusted time series, the seasonal component will be related to average weather conditions. For example, household energy consumption in the UK will increase during the winter caused by colder weather than in the summer. Seasonal adjustment is concerned with identifying a systematic seasonal affect. Using weather or climate data to model energy consumption may reveal a relationship with some temperature variable. Users may be interested in a series that is seasonally adjusted and also weather adjusted. The weather adjustment may be for deviations from average temperature, the average temperature effect forming a part of the seasonal component. For this type of publication weather and climate effects could be an important part of a regular publication. In other situations a one off article exploring weather and climate effects on your time series may be more appropriate rather than including such information in regular publications. This issue is discussed further in section 5.

Does the geographic coverage of the weather and climate data match that of your data? Some of the weather and climate data are available at very detailed geographies and so even if the geographic region you require is not immediately available then it should be possible to construct weather and climate variables for the regions you require. A simple time series model might assume that the weather in a region for a particular time period affects your series in that region and time period. However, you may have reason to believe that weather or climate from different regions or time periods affect your time series. Thinking about what sort of hypothesis you would like to test, where geography plays a part will help to determine what sort of weather and climate data will be appropriate.

Would a bespoke weather or climate data time series weighted by some other variables be appropriate? Many of the time series data in official statistics measure a phenomenon in some geographical area such as the UK. However, the distribution of what is being measured may not be equal across the region. For example, we may expect retail sales to be greater in areas of larger population. Therefore it may be useful to weight weather or climate data together using some other variable. For example, information on population could be used to weight temperature information. Some form of implicit population weighting of weather data is done by the DECC, where data from selected weather stations near to areas of high population are used to create the weather and climate data used in their models; for more information see (DECC, 2011).

Does the frequency of the weather and climate data that you are interested in using in your analysis match that of your data? In the models explored in this guide, the weather and climate data should be of the same frequency as the data you want to model. For example, modelling monthly road accident statistics should have a monthly weather or climate variable. If the frequency of your data and the weather or climate data of interest do not match you may wish to consider aggregating the higher frequency data to the same period as the lower frequency data. Alternatively there are methods of estimating higher frequency data from lower frequency data. Some time series models allow for mixed frequencies within the model. However, these are not discussed in this guide. Contact Time Series Analysis Branch for further information.

How should the weather or climate data be aggregated over time or geography? There are many ways in which weather or climate data could be combined or aggregated to create appropriate variables for your model. For example, for any time period, taking the average maximum daily temperature, the maximum or minimum temperature, or the number of days where the temperature went over or under specified thresholds. Variables may also be aggregated over geographies and you will need to consider what data are available and how the data are to be used in building your models. The practical examples provided in this guide provide some details on derivation of weather or climate variables to include in models.

Are there alternative derivations of weather and climate variables from the available weather data that could be appropriate for your analysis? This is related somewhat to aggregation of weather or climate variables over time or geography, but is more general. For example, we could construct a nice month variable as an indicator variable that takes the value one, if the month was warmer and drier than average and zero otherwise. There are many types of variable that could be constructed, and as has already been stated it is important to construct a hypothesis to test, as this will help you identify what weather and climate data are required.

Does the span match that of your data? In the models presented in this guide weather and climate data are used as exogenous regressors (that is to say we assume that they are not stochastic and follow a deterministic process) and therefore the span of weather or climate data should be at least as long as the span of your own data. In these examples missing observations are not allowed. Some other types of time series models can deal with missing data but these are not discussed here. If you have some missing observations it may be possible to impute values for these. However, you should be aware of the effect this has on your modelling as the imputation introduces additional error that should be accounted for.

What are the relationships (correlation) between the weather and climate variables of interest? It is useful to explore the correlation between weather and climate variables that you wish to consider in your model building. In part this is useful as you may be able to make your model simpler (more parsimonious) by reducing the number of variables, or use this information to construct new variables, that again could be more parsimonious. Some care should be taken to consider the relationships between certain variables. For example, more sunshine hours may be associated with lower than average temperatures in winter months, but higher than average temperatures in summer months.

Back to flow chart