**A Spatio-Temporal Analysis of UK Rainfall Data**

1. **Introduction and Data Description**
   1. Aims

The aim of the project is to analyse the spatio-temporal characteristics of historical UK rainfall data and to employ statistical modelling techniques, such as Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANN), to determine whether the data shows attributes that enables it to be modelled based on its past values such that future values can be accurately forecast.

Time series analysis involves designing a model to enable extrapolation into the future based on the expectation that historic patterns will repeat over time. Several time series modelling methodologies have been developed of which ARIMA is one. ARIMA models have their roots in electrical engineering [1] and were first adapted for analysis of time series using statistical methods by Box and Jenkins in the 1970s [2]. However, there are limitations with ARIMA models, some of which will be discussed including the stationarity requirement and ability to deal with non-linear data. This has led to the emergence of artificial intelligence models such as ANN as powerful alternatives for time series forecasting [3].

* 1. UK Regional Rainfall Data

Monthly rainfall records for the UK were sourced from the UK Met Office [4]. The inspiration for using this data came from an article describing how the data was recently updated through the citizen science project 'Rainfall Rescue' [5]. The data covers the period January 1836 to December 2021 (2232 months) for each of the 10 district regions of the UK as defined by the UK Met Office [6].

Data for each region was downloaded as a .txt file, combined and indexed in an Excel file, and converted to .csv format (“uk\_rainfall\_data.csv”) to enable them to be joined with shapefiles. Shapefiles for the Great Britain district regions were sourced from the CEGE0042 Tutorial data. A further shapefile for Northern Ireland was sourced from OSNI Open Data [7]. These were combined into a single shapefile and then spatially joined with the rainfall data in ArcGIS Pro to create a geodatabase feature layer for importing as a dataframe into R: layer name “uk\_rain\_all\_districts” in file “UK Rainfall.gdb”.

* 1. UK Weather Station Rainfall Data

Monthly UK rainfall data for weather stations sites across the UK was also sourced from the UK Met Office to give a point dataset to complement the areal dataset described in 1.2. The complete record contains monthly data for 37 weather station sites, with records of differing lengths going back as far as 1853 in some cases (see “Station Data.xlsx”). Due to not all the weather station sites having complete records, a subset of sites was identified that had near complete records for a significant period (Sep’64 to Aug’16 (445 months)) and the corresponding rainfall data was segregated for analysis (see worksheet “Selected Combined” in file “Station Data.xlsx”). Gaps in the data for individual months for individual stations were filled to create continuous record and to avoid the issue of NAs when working with the data in R (highlighted yellow in worksheet “rainfall\_by\_station” in file “Station Data.xlsx”). Gaps in month *t* are filled by taking the value for the same month in the previous year (*t-12*) for the same station as a ratio of the sum of the values for all other stations in the previous year (*St-12*) and applying that ratio to the current month’s data for all stations (*St*), as follows: (1.1)

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedHistograms of both the monthly regional data and the monthly weather station point data show non-normal distribution of rainfall values, bounded by zero with a positive skew due to small numbers of months with very high rainfall (Figure 1). Mean monthly rainfall for all regions for the full period of 1836-2021 is 88.55mm, with a standard deviation of 51.33mm. The ‘Scotland W’ region has the highest mean monthly rainfall at 131.78mm, with ‘East Anglia’ the lowest at 51.12mm.

Figure : Histograms of monthly rainfall data for UK regions (left) and UK weather stations (right)

1. **Exploratory Spatio-Temporal Data Analysis**
   1. Spatial and Temporal Characteristics

Summary visualisation of the full UK Regional Rainfall dataset is challenging due to the sheer number of months and due to there being 10 regions. A sample of the data for a shorter period for two of the regions was chosen and plotted as a time series (Figure 2). The data shows considerable variation from month to month, little evidence of a distinct long-term trend, which is consistent with the findings of Lee (2020) [7].

A picture containing graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Chart, line chart, histogram

Description automatically generatedFigure : Plots of monthly rainfall time series for Scotland N and East Anglia regions for Jan 2000 to Dec 2021

Plotting the monthly UK weather station data reveals a similar pattern to the regional data, with significant variation from month-to-month, no obvious long-term trend up or down, and suggestions of seasonality. Plotting the annual averages showed no obvious patterns (Figure 3), indicating there has been no overall change in UK annual rainfall levels on the past 50-60 years, which is consistent with the findings of Jenkins, et al. (2009) [8].

2D, 3D and dynamic scatterplots of the UK weather station point data do not show any clear relationships between average rainfall levels and latitude, longitude, or altitude, other than a hint that rainfall is higher the further west and north.

Figure : Average annual rainfall for all UK weather stations, 1965 to 2015

Text

Description automatically generated with medium confidenceA picture containing text

Description automatically generatedThis relationship is reinforced by examining heatmaps ordered by latitude and longitude, which confirm highest rainfall levels in the north and west (Figure 4). Latitude appears to be a greater factor than longitude. This could be due to the shape of the UK, which has a greater north-south extent than east-west. These heatmaps also show that the spatial variation in rainfall is greater than the temporal, as most of the rows are consistent in colour.

Mapping of the annual average rainfall figures for the weather station point data also confirmed the pattern of wetter in the north and west, drier in the south and east and away from the coasts. Eskdalemuir in Scotland records the highest annual average rainfall (140mm); Cambridge records the lowest (46mm). This is in line with expectations based on knowledge of the south-westerly prevailing weather pattern in the UK, which brings Chart, scatter chart

Description automatically generatedrain in from the Atlantic with the central hills and mountains of England and Scotland creating a rain shadow effect to the east. Choropleths for each month of 1984, which was identified by analysis in Excel to be year with each region most closely matching their annual average rainfall, also clearly shows that the western and northern regions of the UK are the wettest, although ‘N. Ireland’ is relatively drier than nearby ‘Scotland W’. The driest regions are towards the east and the south. Looking just at 1984, there is an indication of seasonality with more rain in the autumn/winter than spring/summer.

Figure : Heatmaps of rainfall for UK weather stations ordered by latitude (left) and longitude (right)

* 1. Spatial and Temporal Autocorrelation

Global Moran’s I was calculated to test the spatial autocorrelation of rainfall levels across the regions of the UK. This had to be done for GB, i.e., excluding N. Ireland, which otherwise created an empty neighbour set as it is not connected to the rest of the UK. A Global Moran’s I value of 0.516 was calculated within a range of -0.593 to 1.018. The moran.test and moran.mc functions both gave p-values <0.05, confirming statistically significant autocorrelation for the regional data, leading to the conclusion that rainfall in one region is more similar in neighbouring regions than those further away.

Figure : Annual average rainfall for UK weather stations, 1965 to 2015

The lowest local spatial autocorrelation values are in the ‘S Wales & England SW’ and ‘England NW & N Wales’ regions. These are wet regions in the west with long borders with dry regions in the east, hence low autocorrelation. The highest local Moran’s I values are seen in ‘East Anglia’, a dry region in the east bordering other dry regions, and ‘Scotland N’, a wet region in the north-west bordering other wet regions. Based on unadjusted p-values, only ‘Scotland W’ and ‘East Anglia’ have significant local Moran’s I, but adjusting the p-values using the Bonferroni method shows no regions with statistically significant local Moran’s I.

Spatial autocorrelation in the UK weather station point data was analysed using a semivariogram. Results show a scattered result but there is an indication that rainfall levels at weather stations that are closer are more similar than those further away. No clear results were seen from the directional variograms, although there are hints of anisotropy in that not all the semivariograms look the same, with the semivariance varying more with distance in the 0o and 135o plots than the 45o and 90o plots. This gives a weak indication that spatial autocorrelation is stronger in the north-south and northwest-southeast directions than in other directions, which is broadly consistent with the findings from the analysis of spatial characteristics.

Temporal autocorrelation within the regional dataset shows it to be generally weak, with one month’s rainfall less strongly correlated to previous month’s rainfall than seen with UK temperatures. The ‘Scotland N’ region shows the highest temporal autocorrelation with a PMCC of 0.306 (Figure 6); Midlands shows lowest PMCC of 0.081 (1 month lag interval).

Chart, scatter chart

Description automatically generatedThe annual data for UK weather stations also shows week temporal autocorrelation (PMCC = 0.121), indicating one year’s rainfall is not significantly related to the previous year’s rainfall.

1. **Methodology and Results**

Figure : PMCC for Scotland N region (one month lag)

* 1. ARIMA

To analyse the data further, the ‘Scotland N’ regional data was selected as this showed the strongest temporal dependency. A shortened version of the ‘Scotland N’ data covering the period 1990 to 2021 is used to make the analysis more manageable. This time series is decomposed into its trend, seasonal and residuals components A picture containing text, people

Description automatically generatedto determine stationarity and then the Box-Jenkins approach to ARIMA modelling is followed to try to fit and test a model for forecasting.

It is expected that the ‘Scotland N’ time series is not stationary due to it containing an element of seasonality. Decomposition of the ‘Scotland N’ time series was done using Seasonal and Trend Decomposition using Loess (STL). Various values for the t.window parameter were tried, with a value of 25 chosen as this resulted in the smallest remainders. The results (Figure 7) show a trend component with no clear pattern, but a clear seasonal component. However, remainders are still high at approximately +/-100mm, compared to the seasonal component, which is +/-50 to 60mm. Confirmation of a clear seasonal component indicates that seasonal differencing will be required for ARIMA. With the first step – Exploratory Data Analysis – already completed, the Box-Jenkins approach moves directly to the next step with analysis of the autocorrelation factor (ACF) and partial autocorrelation factor (PACF) plots and differencing to provide more insight into potential ARIMA model parameters.

Figure : Decomposition of Scotland N regional rainfall data 1990 to 2021 using STL

Timeline

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceThe ACF plot of the ‘Scotland N’ regional data (Figure 8, top) shows a seasonal pattern with positive peaks at lags of 12, 24, and 36 months, and negative peaks at 6, 18, and 30 months. This is the hallmark of a non-stationary time series and strongly suggests seasonal differencing with order 12 is required to make the time series stationary, which is a prerequisite of a linear regression model such as ARIMA (D=1, m=12). This was done and the ACF re-run, producing a plot showing that seasonal differencing has largely removed the seasonal component from the data (Figure 8, bottom) and the data now appears stationary. Significant ACF remain at lags 1 and 12 (seasonal lag 1), whilst the PACF shows a small but significant value at lag 1 and exponential decay in the seasonal lags. This suggests both seasonal and non-seasonal MA components maybe required (q=1, Q=1).

Based on the above, an initial model of ARIMA(0, 0, 1)(0, 1, 1)12 was selected for the parameter estimation and fitting part of the process. The Arima() function in R was used, as recommended by Hyndman and Athanasopoulos [9]. The model gave the following results: log likelihood = -11815.62, Akaike Information Criterion (AICc) = 23637.25. The residuals were checked for any remaining autocorrelations using the checkresiduals() function, which showed that the residuals look like white noise with no significant ACF spikes, except for one at lag 18, which suggests there is some remaining weak autocorrelation and that a better model could be found. The Ljung-Box test gives a p-value of 0.5426, which meets the condition of p >0.05 that suggest the residuals are white noise.

Figure : Scotland N timeseries 1990 to 2021 and associated ACF and PACF plots – Undifferenced (top), with Seasonal Differencing (bottom)

Alternative models were then tested to see if they produced better results based on the log likelihood (looking for a higher value) and AICc (looking for a lower value). No changes were made to the seasonal or non-seasonal differencing components of the models so that the AICc comparison remained valid. The results are summarised in Table 1.

Several of the models tested show very similar results. For example, ARIMA(1,0,0)(0,1,1)12 (fit7.Ar) gave very similar results to the initial model tested. The two models differ solely on having a non-seasonal AR component instead of a non-seasonal MA component. This suggests that there is little impact from the non-seasonal components of the model. Based solely on highest (least negative) log likelihood and lowest AICc value, the ARIMA(1,0,0)(2,1,1)12 model (fit6.Ar) performed the best, although this model still showed just about significant ACF value at lag 18, which suggests not all the autocorrelation has been removed. This model can still be used for forecasting but the correlated residuals may mean that the prediction intervals are not accurate [9]. Compared to the initial model, this model additionally has a non-seasonal AR component instead of an MA component, plus two seasonal AR components.

Table : Summary of ARIMA model fit results

Chart

Description automatically generatedThe best fitting model was used to create a training dataset on 176years of the full 186year ‘Scotland N’ dataset with a view to predicting the remaining 10years of the data and comparing to actuals. The Arima() function uses a one-step ahead approach to forecasting such that the most recently available data is used at each point in time to make the next forecast. The results (Figure 9) show the prediction closely matches the actuals for the first 12-18 months, but then the model breaks down somewhat, particularly around months 30-40 where the model is smooth relative to the noisier actuals. The overall shape of the prediction is reasonably good but seems to have a long-term trend component to the amplitude of the seasonal highs and lows. Comparison was then made with the model found through the auto.arima() function. Running this function on the same Scotland N dataset did not result in a seasonal ARIMA model suggestion, which was somewhat surprising. Instead, an ARIMA(3,1,2) model was found. Whilst the AICc result cannot be directly compared with the results for the models tested per Table 1 due to the presence of a first difference component, analysis of the residuals showed significant ACF values at multiple lags that strongly suggested seasonality is not successfully accounted for in the auto.arima model, and the Ljung-Box p-values are much less than the 0.05 level, meaning this model does not pass the test for it to be considered a suitable model for the ‘Scotland N’ rainfall data. Using the auto.arima model for prediction of the last 10 years of the timeseries yields results that are arguably less accurate than for the model found manually.

Figure : ARIMA model prediction (black) vs. actuals (red) for Scotland N region, 2011 to 2021

* 1. Artificial Neural Networks

Neural Networks are a form of supervised learning where labelled training data used to predict labels for unseen data through adaptive discovery of patterns in the data. Neural networks have been shown to learn from experience and estimate complex functional relationships with high degrees of accuracy, given an appropriate number of non-linear processing units [10]. One type of ANN is the multilayer feed-forward network where one or mode hidden layers of nodes takes weighted inputs from an input layer and provides an output to an output layer consisting of one or more nodes via a non-linear function. This makes them suitable for modelling of non-linear (seasonal) data without the requirement seen in ARIMA of having stationary data before modelling and forecasting. This part of the project aims to test the ability of an ANN to forecast UK rainfall based on the historic rainfall data and to compare results with those of the ARIMA models discussed above.

A multi-input, multi-output ANN was constructed using the nnet() function in R, with the intention of using the lagged values of the timeseries to predict the rainfall for forward months. A training set of 80% of the full 186-year regional rainfall dataset for all ten regions was used to train the ANN. Unfortunately, the results achieved were not in line with expectations based on the analysis of UK temperature data provided in the class tutorial. Regardless of how the parameters of the nnet() function were adjusted, e.g., decay and size (number of nodes in hidden layer), it was not possible to generate a predicted dataset that consisted of anything other than a set of almost constant values for each region. Consequently, predicted values performed very poorly compared to actuals. It seems the data was converging to a mean value or something similar.

An alternative approach was taken using the nnetar() function, as described in Hyndman and Athanasopoulos [10]. This function automatically determines the parameters for the number of lags to use and the number of nodes in the hidden layer. If seasonality is identified in the training data, nnetar() will also add seasonal lags into the model. This is like the auto.arima() function discussed in 3.1.

Graphical user interface

Description automatically generatedThe results using nnetar() certainly looked more encouraging than for nnet(). Using the full 186-year dataset for the ‘Scotland N’ region, nnetar() generated a NNAR(28,10) model, meaning 28 monthly lags of data are input into a hidden layer containing 10 nodes. As with auto.arima, no seasonal component was added to the model. Using this model to generate a prediction of rainfall levels for a future 50 months looked reasonable when plotted, although there are no actuals for this period to use as a comparison to determine accuracy. Due to the length of the timeseries, it is hard to visualise effectively, so a shortened dataset was used containing the last 360 months (30 years) to forecast forward 50 months into the future. This produced a NNAR(18,10) model – 18 monthly lags, no seasonal component, 10 nodes in hidden layer – and the resulting forward prediction looks reasonable when plotted as an extension to the actuals, although the amplitude of seasonal variance looks dampened versus historic actuals (Figure 10). A method to create a prediction for a period that allowed direct comparison to, and plotting against, actuals was not found. So other than a visual inspection for reasonableness, no quantitative analysis of the NNAR model performance has been carried out.

Figure : Actuals (black) and future prediction (blue) from nnetar() for Scotland N timeseries

1. **Discussion and Conclusions**

Spatial and temporal analysis of the UK rainfall time series data shows it to be quite weakly correlated spatially and temporally. The wettest regions are in the north and west of the UK and there is some seasonality with the wettest months in the autumn and winter, but with no apparent long-term trend.

ARIMA requires timeseries to be stationary – problem for weather data which is likely to be inherently seasonal. Seasonal differencing required making selecting suitable ARIMA model more complex. Clearly lots of different models that could have been chosen and comparing results there didn’t seem to be much difference for the metrics being compared (AICc, log-like, etc.). auto.ARIMA came up with something else completely that didn’t seem to make sense from the ACF and PACFs. Results of manual model for predicting rainfall appeared better than for auto.ARIMA (is there a way of quantifying this – RMSE?)

References

[1] Wiener, N. (1949). Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications. Technology Press of the Massachusetts Institute of Technology

[2] Box, G.E.P. & Jenkins, G. M. (1970). Time series analysis : forecasting and control. Holden-Day

[3] Mitrea, C. A., Lee, C. K. M., Wu, Z. (2009). A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. International Journal of Engineering Business Management, Vol. 1, No. 2, p 19-24

[3] Met Office (2022). UK and regional series. Available at https://www.metoffice.gov.uk/ research/climate/maps-and-data/uk-and-regional-series. Accessed 04/04/2022

[4] Hawkins, E., Burt, S., McCarthy, M., Murphy, C., Ross, C., Baldock, M., et al (2022) Millions of historical monthly rainfall observations taken in the UK and Ireland rescued by citizen scientists. Geoscience Data Journal, 00, 1– 16. Available from: https://doi.org/10.1002/gdj3.157

[5] Met Office (2022). UK climate districts map. Available at https://www.metoffice.gov.uk/research/climate/maps-and-data/about/districts-map. Accessed 04/04/2022

[6] Open Data NI (2022). OSNI Open Data – 50K Boundaries – NI Outline. Available at https://www.opendatani.gov.uk/dataset/osni-open-data-50k-boundaries-ni-outline. Accessed 04/04/2022

[7] Lee, E.M. (2020). Statistical analysis of long-term trends in UK effective rainfall: implications for deep-seated landsliding. Quarterly Journal of Engineering Geology and Hydrogeology, 53, 587-597

[8] Jenkins, G.J., Perry, M.C. & Prior, M.J. (2009). The Climate of the United Kingdom and Recent Trends. Met Office Hadley Centre, Exeter

[9] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2

[10] Zhang, P.G. & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, 160, p 501-514