ENEN90032 - Environmental Analysis Tools Assignment 1

Dongryeol Ryu, Lilangi Wijesinghe and Manish K. Patel

20 August 2022

Submit an electronic copy (in PDF format) to the *Turnitin* menu of the subject LMS by 9:00pm on Sunday 4 September 2022. Make sure you meet the Infrastructure Engineering submission requirements (include the ENEN90032 coversheet with signatures of team members). Include appropriate graphs and tables in your solution. The report should contain no more than **2500 words** excluding figures and figure/table captions. State contributions of your group members to the assignment explicitly in the last section of your report. The statement of contributions section will not be include in the word count. Submit your Python codes and input data files (*COMPRESSED!*) via your assignment group folder (/MA1_GXX/Files; XX is your group number). Your Python codes should be in Jupyter Notebook format with proper comments and executable section by section as in the Tutorial material. Submit a consolidated Jupyter Notebook file that contains modular codes for all questions.

- Place all the necessary input data files in the same folder with the Python codes (or in the subfolders referenced in the Python codes) so that I can run all the sub-sections on my computer without any modification
- Add proper comments to explain the role of each section and main steps within it
- Your Python code is expected to produce figures and numeric outputs identical to the ones used in the report

For the hypothesis test problems, you must explain rationale behind the choice of the null hypothesis, test statistic, alternative hypothesis and conclusions. 20%-30% of the marks for each section will be given based on the report quality and quality of figures and tables whenever they are required. Figures should be properly labeled and self explanatory. You are expected to work TOGETHER for all

questions, thus issues arising from splitting assignment questions between the assignment group members will not be assisted with.

In summary, submission includes (group submission!):

1. A report ($\leq$ 2500 words) that includes ENEN90032 coversheet and the statement of contributions section via the *Turnitin* menu of the subject LMS
2. Compressed Jupyter Notebook format codes (a consolidated file) and input data deposited in /MA1_GXX/Files folder

1 Exploratory Data Analysis - Meteorological Datasets (15 marks)

Go to the *Climate Data Online*[1] of Bureau of Meteorology and choose a weather station in Perth, Western Australia (or surrounding area with six-digit station number staring with *009XXX*), one in Darwin, Northern Territory (or surrounding area with *014XXX*) and a station in Melbourne, Victoria (station number staring with *08XXXX*). Download daily rainfall data of the stations collected in a year between 2011 and 2021 inclusive (the selected year for the Darwin and the Melbourne stations should be identical). Missing values in the selected year should be fewer than 10. For the rainfall data analysis, we will be using wet-day daily rainfall data, which excludes *zero-rainfall events* and the values *lower than the detection limit (assume that the detection limit is 0.25 mm).*

1. Make a table that summarizes the location (sample mean, median and trimean), spread (sample standard deviation, IQR and median absolute deviation) and symmetry (sample skewness and Yule-Kendall index) of the datasets in the cities. Can you infer skewness of the datasets by comparing the mean with the median? Based on the shape of the distribution (refer to the figures produced in the next question), discuss the robustness of the summary statistics calculated above.
2. For the wet-day daily rainfall data, fit i) a Gaussian, ii) a gamma, and iii) a Weibull[2] distribution functions to the dataset and compare the fitted distribution models with the data distribution. For graphical representation of the probability density of data, use the Gaussian kernel estimates that produce a smoothed curve for the probability density. Also, compare its empirical cumulative distribution with fitted CDFs (Gaussian, gamma, and

[1] `http://www.bom.gov.au/climate/data/`

[2] `https://en.wikipedia.org/wiki/Weibull_distribution`

Weibull) and make a Q-Q plot for evaluation. Judge which model fits your data best based on the graphical examinations.

3. Suggest a probability density model other than the above three models, which closely fits the wet-day daily rainfall data from Perth, Darwin and Melbourne, respectively, and compare the performance of your choice with the results in the previous question. You can use any existing models with 1-to-3 parameters. You may refer to published research articles, e.g., Ye *et al.* (2018)[3] if needed.

4. For the above fits to the rainfall data (including your suggested model), calculate the log-likelihood values of the fits and quantitatively prove your judgement above.

2 Daily Max Temperature at Melbourne and Essendon Airports (10 marks)

In this question, you are testing if the daily maximum temperature values measured in the Melbourne Airport (Tullamarine) and the Essendon Airport are significantly different from each other to justify maintaining both weather stations. Go to the Climate Data Online of BoM used in the previous question and download daily maximum temperature of 2021 (365 samples each location) at the stations #086282 (Melbourne Airport) and #086038 (Essendon Airport). Choose a method from the Hypothesis Test section (e.g., one-sample test, two-sample test, confident-interval based method, p-value based method, etc.) and suggestion a conclusion if the daily temperature values measured at these two locations show difference with statistical significance at 95% and 99% confidence levels (5% or 1% significance levels for p-value based method). Show/describe the procedure, including the test statistic and null-distribution, clearly in both your report and the Jupyter notebook file.

3 Newcomb-Michelson Velocity of Light Experiments (10 marks)

Simon Newcomb of the Nautical Almanac Office (NAO), U.S., published the velocity of light [Newcomb, 1883][4] based on a series of experiments he conducted with Albert Michelson until 1882. The dataset 'NewcombLight.txt' contains 66 samples (time in seconds taken for light to travel 7442 meters at sea level) Newcomb

[3] `https://doi.org/10.5194/hess-22-6519-2018`

[4] `http://vigo.ime.unicamp.br/~fismat/newcomb.pdf`

collected in 1882. Conduct the t-test and the bootstrap based one-sample tests and provide the population mean of light velocity (in m/s) with your choice of a confidence level. Do the estimates include the widely known speed of light as in HERE[5]? Do the estimates from the t-test and the bootstrap show any systematic difference? If so, provide possible reasons based on the sampling distributions used by the two approaches.

4 Space Shuttle O-Ring Failures (10 marks)

On 27 January 1986, the night before the space shuttle Challenger exploded, engineers at the company that built the shuttle warned NASA scientists that the shuttle should not be launched because of predicted cold weather. Fuel seal problems, which had been encountered in earlier flights, were suspected being associated with low temperatures. It was argued, however, that the evidence was inconclusive. The decision was made to launch, even though the temperature at launch time was 29 °F (~ -1.67 °C).

The dataset 'O_Ring_Data.XLS' summarizes the number of O-ring incidents on 24 space shuttle flights prior to the Challenger disaster. Launch temperature was below 65 °F for data labeled 'COOL' and above 65 °F for data labeled 'WARM'. Conduct a permutation test if the number of O-ring incidents was associated with the temperature using 99% confidence interval with your choice of one-sided or two-sided test options. Use 10,000 permutations to draw conclusion. Justify your choice and show your null distribution as a histogram with a test statistic marked on it. Make your final suggestion about the launch of the space shuttle on the day of accident based on the quantitative evidence that supports your suggestions.

5 Cloud Seeding Experiment (15 marks)

The dataset 'Cloud_Seeding_Case_Study.XLS' contains data collected in southern Florida between 1968 and 1972 to test a hypothesis that massive injection of silver iodide into cumulus clouds can lead to increased rainfall (J. Simpson, and J. Eden, "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification," *Technometrics* 17 (1975)). An airplane flew for 52 days in total, however, silver iodide was injected on randomly chosen 26 days. The pilot was not aware of whether on any particular day the cloud seed was loaded or not to prevent biases. The rainfall was measured by radar as the total rain volume falling from the cloud base following the airplane seeding.

[5] `https://en.wikipedia.org/wiki/Speed_of_light`

1. Using a parametric method, conduct a test if the cloud seeding made a significant impact on rainfall using both 95% and 99% confidence intervals. Choose between one-sided and two-sided tests and justify your choice.

2. Repeat the above test now using a permutation test. Use 10,000 permutations to draw your conclusion and show your resampled data in the histogram. Compare your results with those from the parametric test above and explain the differences identified based on the pros and cons of the two methods.

3. You may have noticed that the rainfall data in the cloud seeding experiment are highly skewed. Transform the rainfall values using a logarithm function and repeat the parametric test under the same conditions used for the Question 4.1 above. Does the transformation change your conclusion? If so, discuss about the difference and the implications of the results to the need of data-transformation when the data is highly asymmetric.

6 Exploratory Data Analysis and Linear Regression (20 marks)

Nutrient/sediment concentrations vs. stream discharge relationships have been widely used as a clue to explore hydro-chemical processes that control runoff chemistry. Here we examine sediment concentration vs. stream discharge relationships using linear regression. This question investigates the correlation between instantaneous streamflow and the Total Kjeldahl Nitrogen concentration (TKN) collected from the site "222101, Curdies River at Curdie" located in Otway Coast of Victoria. The data 'Q_TKN_data.csv' contains three columns: the monitoring date (column 1), catchment-averaged streamflow (mm/d, column 2) and TKN (mg/L, column 3).

1. Calculate the Pearson correlation coefficient and Spearman's rank correlation coefficient for the paired Q vs.TKN data. Then, calculate the same correlation coefficients for the paired natural log(Q) and natural log(TKN) i.e. logarithm with a base e. What do these values tell you about the relationship between TKN and Q? Suggest which paired data (raw data pair vs. log(data) pair) is more suitable for constructing linear regression, and justify your selection. In the subsequent questions, Q vs. TKN means their relationship based on your chosen pair.

2. Based on your selection of the paired data, plot Q vs. TKN concentrations and fit a simple linear regression: i) report the regression parameters and

the goodness-of-fit for the regression; ii) use the linear regression developed, predict the TKN concentration expected when discharge reaches 2 mm/d.

3. Calculate the 95% confidence intervals for i) conditional mean and ii) prediction. Construct a figure showing the linear model and the confidence intervals with the observed data values and discuss the difference between the confidence intervals for conditional mean and prediction (e.g., how to interpret and use the intervals?).

4. What is the pattern of the residuals of your developed Q∼TKN model? Provide your assessment of the residuals. Assessment of autocorrelation (serial correlation) in residuals is needed.

5. Based on the plot you created in Question 6.3, check how much fraction of the observed data values actually falls within the 95% prediction confidence interval (e.g., you can create a `for` loop in Python to check if individual observation falling within 95% CI). According to the results and the pattern/distribution of residuals, do you recommend the application of this linear model for predicting further TKN concentrations? Did you find any specific range of Q where your model struggles to predict TKN (for this last question, compare the predicted TKN with observed TKN values in the original (raw data) space in case you built your model in the log-transformed space)?

7 Atmospheric CO_2 Concentration during Global Forced Confinement by COVID-19 (20 marks)

Global forced lockdowns caused by fast spreading COVID-19[6] since late January 2020 reportedly reduced global CO_2 emission. A report[7] published in the early stage of COVID-19 estimates the reduction in CO_2 emission as high as 17%. Its follow up study[8] published in March 2021 reports changes in CO_2 emission in the post-COVID-19 era. In this section, we examine *if the atmospheric* CO_2 *concentration was lower than the level it would have been without COVID-19* during the pandemic period including the peak forced confinement period, from April 2020 to March 2021.

[6] `https://ourworldindata.org/grapher/covid-stringency-index?time=2021-07-19`
[7] `https://www.nature.com/articles/s41558-020-0797-x`
[8] `https://doi.org/10.1038/s41558-021-01001-0`

To examine the atmospheric CO_2 in April 2020–March 2021, we use the monthly CO_2 data[9] archived in the Scripps Institution of Oceanography of the U.S. If you are interested in the history of the long-term atmospheric CO_2 measurement and its current status, check this link[10]. Download the monthly CO_2 data HERE[11] and use the unadjusted values in the *5th column* this time. We will be using the monthly CO_2 concentration values from *April 1995*.

1. Monthly timeseries of atmospheric CO_2 features steady increase from the beginning of monitoring with strong seasonal fluctuations. Consequently, anomalous behavior of CO_2 needs to be examined after removing the background trend and seasonal fluctuations. Construct CO_2 concentration anomaly by removing long-term trend and seasonality before the total dataset (April 1995 – March 2021) is separated into *control* (April 1995–March 2020) and *treatment* (April 2020 – March 2021). For example, you can remove the seasonality of data by subtracting the mean (arithmetic) of all CO_2 values of month X (in 1995–2021) from individual monthly data of month X. Use a polynomial function to remove the long-term trend. Describe your method and summarize results as time series of anomaly (residual) CO_2 concentrations and Gaussian kernel densities (or histograms) of the control and treatment samples.

2. Now you conduct a hypothesis test to examine *if the atmospheric* CO_2 *concentration was lower than the level it would have been without COVID-19* during the pandemic period. Choose a method most suitable for the problem (parametric vs. non-parametric, confidence interval based vs. p-value based, one-sided vs. two-sided, etc.) and draw conclusion based on your test. Provide the null and alternative hypotheses for the test. Provide key metrics, numeric or/and graphical, that supports your test.

3. What is your assessment of the atmospheric CO_2 concentration in April 2020–March 2021 based on the test statistics in the previous question? Are they within the range of your intuitive expectations? There exist a large number of academic and general articles about the expectations and interpretations around the observed atmospheric CO_2 published online in 2020-2022. Provide your assessment and interpretation of the test statistic supported by your choice of relevant articles.

[9] `https://scrippsco2.ucsd.edu/data/atmospheric_co2/primary_mlo_co2_record.html`

[10] `https://keelingcurve.ucsd.edu`

[11] `https://scrippsco2.ucsd.edu/assets/data/atmospheric/stations/in_situ_co2/monthly/monthly_in_situ_co2_mlo.csv`