IE5374 Fall 2022 - Project 2

The objective of Project 2 is to exercise the skills you've learned for Time Series Analysis and Forecasting.

Task 1

The following link provides human activity data for 15 subjects (Under the DataSet - RealWorld (HAR) (2016) link). You can access a subject's time series data by clicking on that subject. For this project, consider accelerometer data on the chest for 3 subjects of your choosing for walking, running, climbing up, and climbing down.

Complete the following:

1. Select a sample size of 1024 points (from 1000 to 2023) for each of your 3 subjects.
2. Apply a natural visibility graph (NVG) and horizontal visibility graph (HVG) to each of the aforementioned data. You have 3 subjects, 3 accelerometer directions, and 4 activities for each, so you should end up with 72 graphs. **You DO NOT have to display the graphs!!!**
3. For each graph, compute the average degree, network diameter, and average path length
4. Tabulate all the results (Sample Table Output shown below)
5. Generate scatter plots of average degree vs network diameter and color the points according to walking and running (do this for each accelerometer signal and each method (NVG and HVG))
6. Generate scatter plots of average degree vs network diameter and color the points according to climbing up and climbing down (do this for each accelerometer signal and each method (NVG and HVG))

Method	Subject	Accelerometer axis	Activity	Average Degree	Network Diameter	Average Path Length
HVG or NVG	1 to 15	X, Y, or Z	Walking, running, climbing up, or climbing down			

Task 2

Google has made available some mobility data around the world since the COVID-19 pandemic began. You can find their datasets here. For this task, you will be forecasting using data from the King County. There are 2 ways to get these data:

- Download the Global file (not recommended, it’s huge)
- Download the folder of all Regional files. There should be 3 years of CSV files for each region, you should locate and grab the 3 corresponding to “US”

Once you have the proper data files, you should be able to isolate just the data from King county via the `sub_region_#` columns. In this activity, we will be concerned with describing and forecasting the time series concerning “Residential”, “Work”, and “Grocery and Pharmacy”

1. Put together your entire time series using all the data from 2020-2022. You should end up with 1 dataframe that contains all the data points.
2. Trim down your time series to remove the months before April 2020. This will remove the very early pandemic and the pre-pandemic conditions.
3. For each of the 3 time series, perform an additive Time Series Decomposition and plot the results. You should show the trend, seasonality, and remainder in your plots.
4. For each time series, build a forecasting model using Exponential Smoothing (ES). You should test out at least 2 different ES models and use forecast evaluation metrics (e.g. MAE, RMSE) to demonstrate why you chose your best ES model
5. For each time series, build a forecasting model using ARIMA. You must show why you chose your ARIMA model.
6. Compare your best ES and best ARIMA models for each time series using forecast evaluation metrics. Show which model is best in each case.
7. Using your best model, forecast the rest of 2022 for each time series. Show these forecasts by plotting the past data points in 1 color and the future data points in a second color.

Submission Format

1. Submit all your solutions in an iPython notebook (`.ipynb`)
2. Make sure you show your code AND results
3. This is a group assignment; only one member must submit per group
4. Submit your solutions by December 5th @ 12pm PT