Time Series End term

Write the equations for all the models used. Give justifications for each step. Give proper interpretation of results, and conclusions.

The data consists of 2 annual series, each related to some (undisclosed) tourism activity. Tourism activities may include inbound tourism numbers to one country from another country, visitor nights in a particular country, tourism expenditure, etc. The series may differ in the order of magnitude of values.

1. Plot all the series (an advanced data visualization tool is recommended) - what type of components are visible? Are the series similar or different? Check for problems such as missing values and possible errors.
2. Partition the series into training and validation, so that the last 4 years are in the validation period for each series. What is the logic of such a partitioning? What is the disadvantage?
3. Generate naive forecasts for all series for the validation period. For each series, create forecasts with horizons of 1,2,3, and 4 years ahead (F_{t+1}, F_{t+2}, F_{t+3}, and F_{t+4}).
4. For each series, compute MAPE of the naive forecasts once for the training period and once for the validation period.
5. The performance measure used in the competition is Mean Absolute Scaled Error (MASE). Explain the advantage of MASE and compute the training and validation MASE for the naïve forecasts.
6. Create a scatter plot of the MAPE pairs, with the training MAPE on the x-axis and the validation MAPE on the y-axis. Create a similar scatter plot for the MASE pairs. Now examine both plots. What do we learn? How does performance differ between the training and validation periods? How does performance range across series?
7. For forecasting, first compare the three methods and then use an ensemble of the three methods:
   • Naive forecasts multiplied by a constant trend (global/local trend: "globally tourism has grown "at a rate of 6% annually.")
   • Linear regression
   • Polynomial regression
   • Exponentially-weighted linear regression
   (a) Write the exact formula used for generating the first method, in the form $F_{t+k} = \ldots$ (k = 1, 2, 3, 4)
   (b) What is the rational behind multiplying the naive forecasts by a constant? (Hint: think empirical and domain knowledge)
   (c) What should be the dependent variable and the predictors in a linear and polynomial regression model for this data? Explain.
   (d) Fit the regression models to both the series and compute forecast errors for the validation period.
8. If you are to consider exponential smoothing, what particular type(s) of exponential smoothing are reasonable candidates? Discuss the results of ES model that you considered.
9. Can you suggest methods or an approach that would lead to easier automation of the ensemble step?