

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

Sept 2023

AUDIT TRAIL

**Subject CP2 - Modelling Practice
Core Practices**

Paper One

Objective

Our client is a multinational company that is looking for some analysis to be done on global temperature changes. As they deal with refrigeration, climate changes have an impact on their costs, and the purpose of this model is to determine the possible increase in costs the company may experience over the next decade.

This is done by analysing the temperature changes over the past years from the given data, fitting a linear regression model to the data and hence predicting future temperatures by identifying a trend. This will help assess how future refrigeration costs are expected to change with the predicted changes in temperature.

Data

Data was provided by my manager. (The source is the National Centres for Environmental Information (NCEI)). It consists of two tables: The **Stations** tab holds information on each weather station our client has operations in (location, elevation and country/continent). The **Data** tab holds average monthly temperature in degrees celsius for each year for each weather station. The years range from 1961 to 2010.

Data Analysis and Checks

Stations tab

On the Stations tab, a number of checks were done:

- A check was done in column H to ensure that each weather station featured in the Data sheet. This is done with a MATCH function. No errors were found, so there is temperature data for all 246 weather stations in the **Data** sheet list
- In the area to the right of the data, max, min and average is taken of latitude, longitude and elevation.
 - Latitude (max<90, min>-90) and longitude (max<180, min>180) look reasonable.
 - Elevation doesn't look right. Minimum of -22 is plausible - there are places on earth (for example Netherlands, Israel) which are below sea level. But max of 9999 is wrong, and is clearly a default value. This will need to be changed.
- Below that, a table is created (using COUNTIFS) of the number of stations per continent. These appear to be reasonably well spread around the world, with more being in Europe and North America. This feels reasonable as it is where GDP is greatest, and so is likely to be where more clients needing refrigeration would be found.

Stations Adj Tab

This tab is a copy (by reference) of the data in the **Stations** tab. The only change is that where elevation in column D = 9999, it is set to zero instead using an IF function. Here, the max elevation (3700m) and average elevation (470m - close to the 500m input) look reasonable.

Data Tab

On this tab, checks are done on the temperature data:

- In column O, average temperature is calculated for each year across all months. This is then used in columns R-T
- In columns R to T, an average across all stations per year was taken, as well as a count of the number of observations for each year. The number of observations start at 245, but move to 246 by 1965 and stay there - clearly there was one station that only started operating in 1965. No change needs to be made for this as we're only working in averages.

A graph is produced for the average temperature each year - this shows a bit of annual variation, but a general trend from 13 degrees in early 1960s moving up to 14 degrees around 2010 can be seen.

- In columns X to Y, a table of min, max and average temperature for each month is created using MIN, MAX and AVERAGE applied to the monthly columns (C to N). This shows that minimum temperatures clearly have a default applied of -99.99. These values will clearly have to be ignored. Average temperatures rise in the middle of the year and drop over the start/end. This is expected due to the majority of weather stations being in North America and Europe: they'd have summer over the middle of the year, and winter from Nov to Feb. Max temperatures have a similar, but far less pronounced pattern.
- A chart is produced of the monthly min, max and average, which displays the results discussed in the previous point.

Data Adj tab

The data from the **Data** tab is copied across by reference. The only changes are to replace the default values of -99.99 by NA. Having a non-numeric reference will remove it from the average calculations that we will be doing. I considered doing something more complicated, for example by taking the average of the value before and after the default value, but there are some cases where there are two or three defaults in a row, and this would be awkward to achieve in Excel. Given that we are taking averages anyway, the approach to ignore the missing values seemed the best course of action.

The calculations and checks are copied as well, and the temperature graph shows more reasonable values for minimum temperatures - again, this is affected by the bias towards the Northern hemisphere, and shows a marked increase in summer and a decrease in winter.

Assumptions

The following assumptions have been made:

- After the adjustments described above, the data is fit for purpose and accurate.

- Summer months apply consistently as April to September across the whole northern hemisphere, irrespective of latitude
- No more recent suitable data is available after 2010
- No event or other changes have taken place to invalidate the trend obtained from the data up to 2010.
- No change in business conditions that affect refrigeration costs will take place between 2020 and 2030

Inputs

This worksheet provides a list of all parameters to support the calculations. This includes inputs to split locations by latitude above and below a certain point (set to zero), and by elevation (set to 500).

Also here is information about costs in 2020 and the constant parameter for the formula used to determine the impact of temperature on refrigeration costs.

Method

Each sheet is dealt with in turn.

Data Prep

This sheet does the analysis of the data and prepares the summary from which the later graph and linear regression will be done.

Columns A to N are copied by reference from the **Data Adj** sheet.

Column P contains the Hemisphere category - whether this is North (latitude >0) or South (latitude < 0) based on a lookup of the weather station ID on the **Stations Adj** tab.

Column Q contains the elevation category - whether the station is above or below the input elevation parameter of 500m. The result is returned as High or Low based on a lookup on the **Stations Adj** tab.

Column R contains the average temperature across the year - taking the average of columns C to N.

Column S contains the average temperature for Winter. This is the average of April to September for Southern Hemisphere stations (based on column P), and the average of the other months (Jan to March and October to December) for Northern Hemisphere stations.

Column T contains the average for Summer - this is the same as for Winter, but reversed.

Column U picks up the continent based on a lookup of Weather station ID from the **Stations Adj** tab

Analysis

This tab calculates average temperatures for each year for each category.

The Averageifs formulas is used throughout to calculate the average across all data points that meet the provided criteria. Unless otherwise specified, the average of column R on the Data Prep sheet is taken.

Column A has the year, from 1961 to 2010

Column B calculates the overall average temperature across all stations for each year.

Column C and D calculate the average temperature across all Northern Hemisphere and Southern Hemisphere stations respectively for each year (using the category in column P of the Data Prep sheet)

Column E and F calculate the average temperature for all Low and High elevation stations respectively for each year (using the category in column Q of the Data Prep tab).

Column G and H calculate the average Winter and Summer temperature for each year (taking the average from column S and T on the Data Prep tab respectively).

Columns I to N calculate the average temperature for each continent each year (using the continent in the Data Prep tab column U).

Charts

This tab contains the following four line charts showing movement of temperature over the years from 1961 to 2010, broken down as follows:

- Weather stations north and south of the equator, as well as an overall average. This also shows the bias to stations in the north: the overall average is much closer to the north line. The north average is much lower than the south, and shows slightly more volatility (such as from 1995-1999).
- Weather stations higher and lower than 500m elevation. This has a minimal impact on average temperature, with the three lines very closely banded. While higher elevation should (and does) have lower average temperatures, it is likely that most weather stations would not be at the really high altitudes which would have a more dramatic impact on temperature.
- Average temperature for summer and winter. Summer is much higher than winter, as expected. Summer temperatures also appear to be smoother over time than winter.

- Average temperature by continent. Africa and South America are well clear above 20°, with Europe notably lower than the rest below 10°. It appears that the continents with higher temperature have lower volatility.

Model

This tab uses a linear regression model to predict future summer temperatures, and uses those to gauge the impact on refrigeration cost.

At the top of the sheet, in columns B and C, the LINEST formula is used to do a linear regression. This formula is an array formula, so care must be taken when editing this to select both cells B6 and C6, and use Ctrl-Shift-Enter to enter it. The formula takes Y values as the average summer temperatures from column H in the **Analysis** tab, and X values as the year.

The linear regression formula is $y=mx + c$
 m is the coefficient - in cell B6. c is the constant, in cell C6.

Below this, the actual and predicted summer temperatures are calculated. From 1961 to 2010 the formula just links to column H on the **Analysis** tab. From 2011, the formula above is used to calculate the predicted temperatures. The cell for 2011 is highlighted to avoid formulas being copied down.

In column C, the expenses are calculated, using the formula

$$E(t) = E(t-1) * e^{(k * (T(t) - T(t-1)))}$$

Where E(t) is the expense in year t (the expense in 2020 is set to \$250m from the inputs tab)

K is a constant (from the inputs tab)

T(t) is the average temperature in year t.

A line chart is produced showing the average summer temperature from 1961 to 2030, with the last twenty years being the predicted values (hence they're a straight line).

In columns F-H, the average temperature for 2006-2010 (in H13) and 2026-2030 (in H14) are calculated.

The average costs for 2026-2030 are also calculated just below this.

Then, the increase in temperature over that period is calculated as 1.87%