

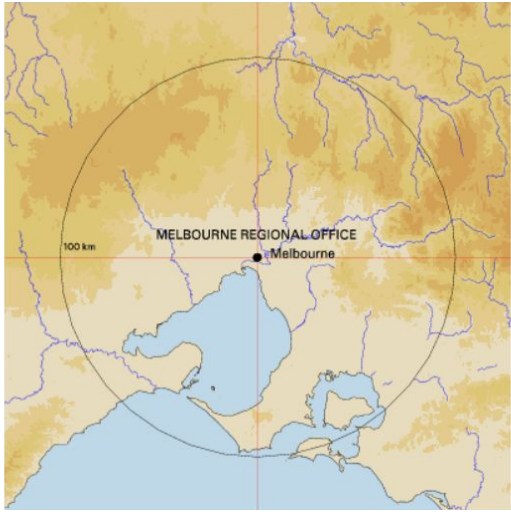
Mini Project Proposal

Course: MATH-342, Time Series (Spring 2017-18)

Wong Wen Yan, Zhan Fengyue

wen.wong@epfl.ch, fengyue.zhan@epfl.ch

General Information

Name of dataset	Daily Maximum Temperature at Melbourne Regional Office
Source of dataset	Bureau of Meteorology, Australian Government
Link	http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_display_type=dailyDataFile&p_stn_num=086071&p_nccObsCode=122&p_startYear=2003&p_c=-1481645376
Description	The daily maximum air temperature is nominally recorded at 9 am local clock time. It is the highest temperature for the 24 hours leading up to the observation, and is recorded as the maximum temperature for the previous day.
Station number	86071
Year site opened	1908
Year site closed	06 Jan 2015
Latitude (decimal degrees, south negative)	-37.81
Longitude	144.97
Number of instances in dataset	58444
Location on Map	

Information about dataset

Column	Attribute	Data type
1	Product code	String
2	Bureau of Meteorology station number	Integer
3	Year	Integer
4	Month	Integer
5	Day	Integer
6	Daily maximum temperature (degree Celcius)	Float
7	Period over which daily maximum temperature was measured (days)	Integer
8	Quality of daily maximum temperature	String ('Y' or 'N')

Remarks

For purposes of statistical analyses, several columns are removed from the dataset, namely,

Columns	Reason
1, 2	Values are constant across all instances in the dataset, and thus do not hold relevance
7	This is an external variable, which is not useful for the time series models studied in this course

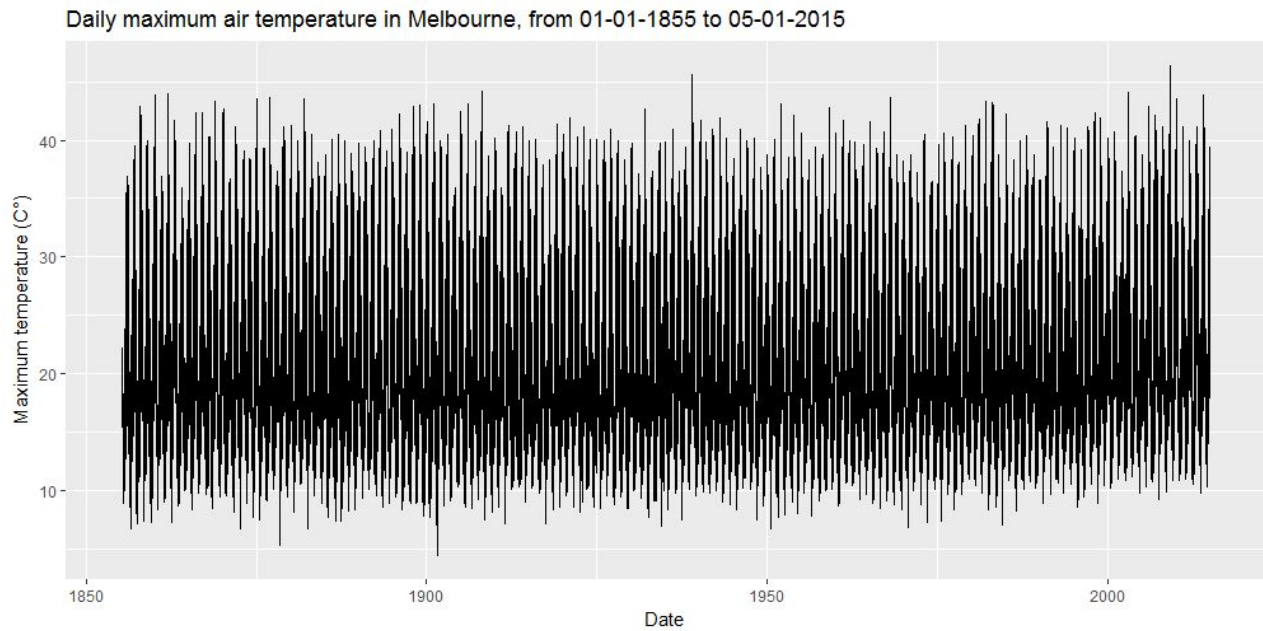
According to descriptions from the data provider, column 8 indicates the Quality Flag of an instance with the following meanings:

‘Y’: completed quality control and acceptable

‘N’: not yet completed quality control process or date uncertain, quality not assured

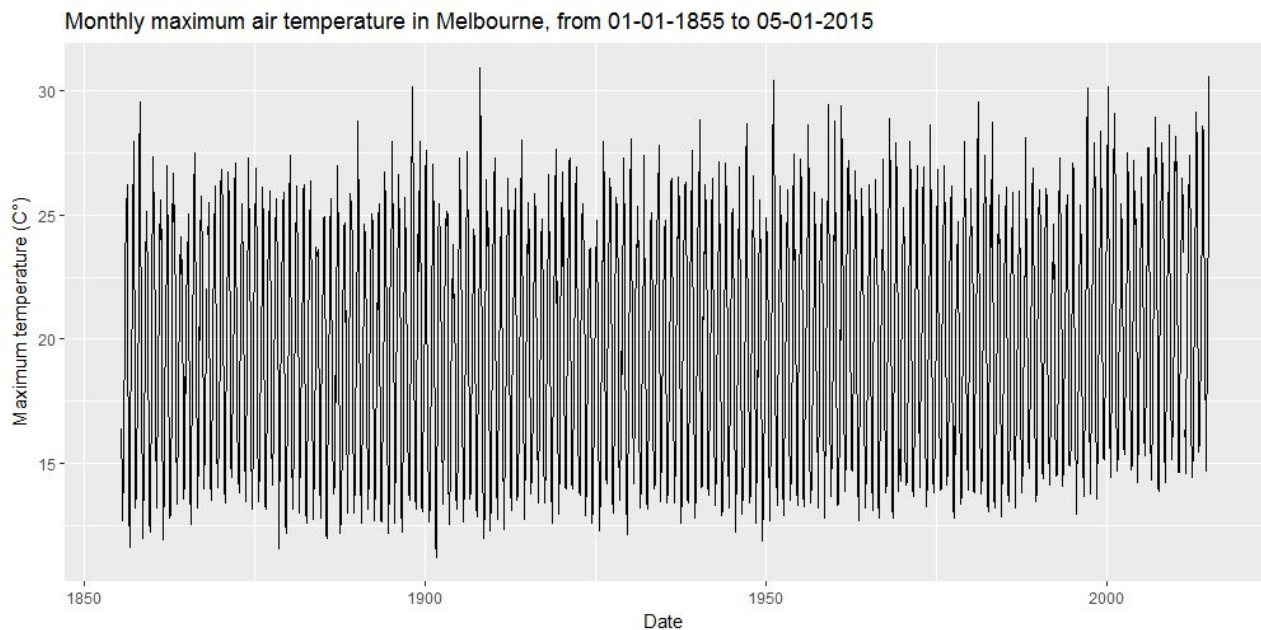
To visualize the portion of ‘low quality instances’ in our dataset, we have made several plots in R.

Visualization of Daily Maximum Temperature series (without low quality instances)

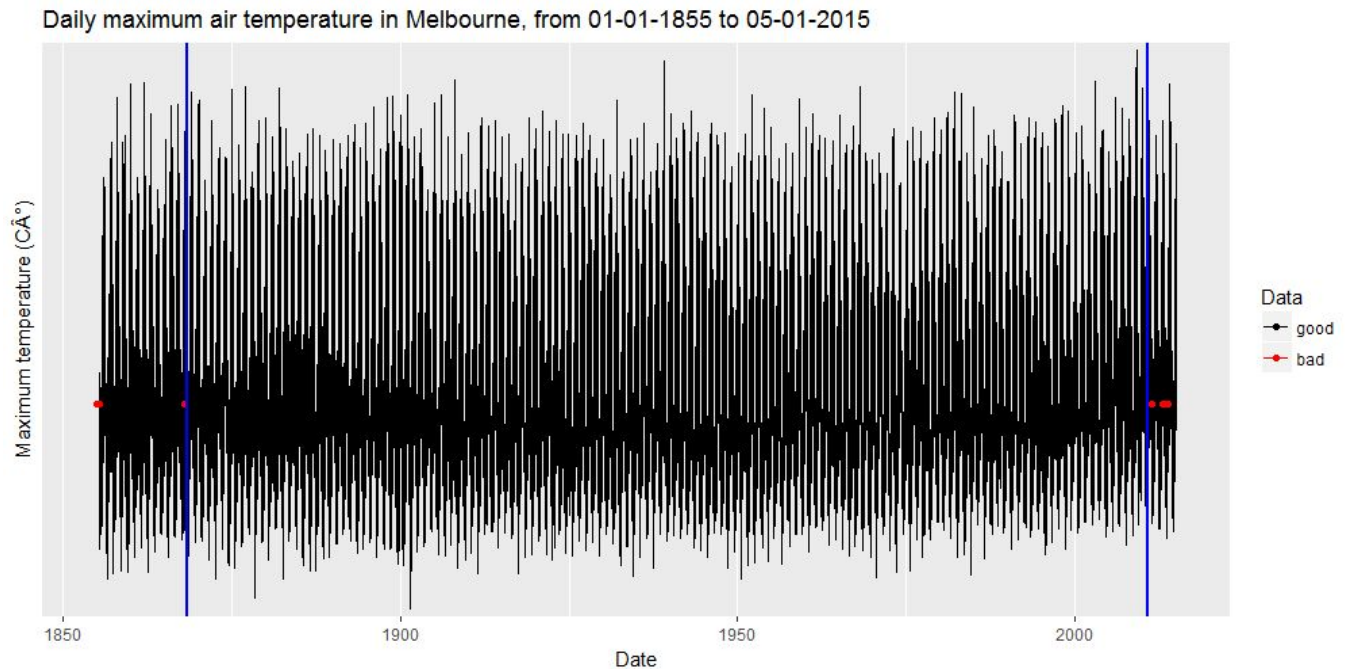


Due to the large number of instances, it is difficult to visualize the entire daily series with limited pixel resolution in R. Thus, we have produced another plot by aggregating daily maximum temperatures of a month into a single point (through averaging).

Visualization of Monthly Maximum Temperature series (without low quality instances)



Visualization of Daily Maximum Temperature series (with low quality instances)



There are a total of 216 bad instances in the entire series, which is approximately 0.3% of all available instances. It can be seen that low quality instances are most prevalent at the beginning and end of the dataset. If we measure the longest segment of continuous data points without low quality data (ie. number of instances between the two blue lines), this is approximately 89.05% of the entire dataset. Since our dataset is large (58444 instances), we have decided that the easiest solution would be to remove all instances at the beginning (segment before first blue line, 4744 instances) and at the end (segment after second blue line, 1657 instances), with remaining 52043 instances for further analyses.

Proposed analysis:

Goal:

1. Construct a forecasting model, try to understand main seasonal cycles within the temperature series
2. Recover data with bad quality by estimating actual value using a trained forecasting model

Analysis (tentative):

1. Subtract mean and seasonality of series (using periodogram)
2. Check if residuals is Gaussian i.i.d
3. If no, model the residuals with ARIMA, GARCH, etc
4. After fitting residuals to different models, use an averaged forecast to obtain the best prediction.