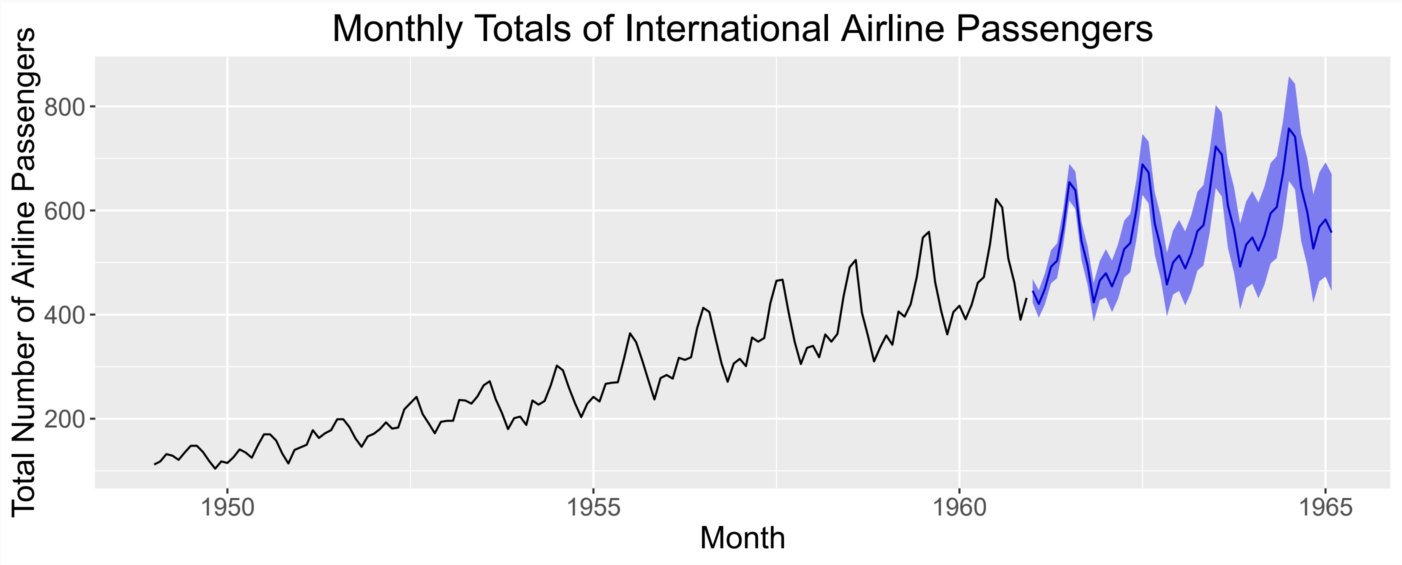
Time Series Interpolation Algorithms:

An Application to Real-World Data

Melissa Van Bussel
Trent University
Canadian Statistics Student Conference,
May 25th, 2019

Recall...

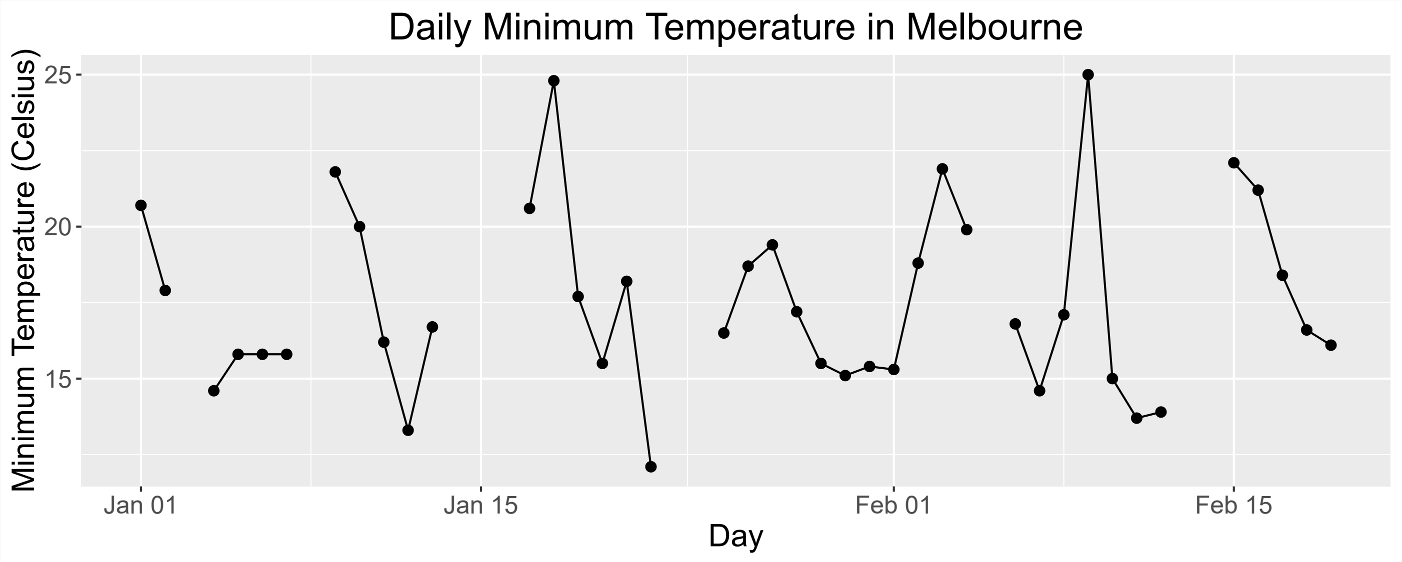
A **time series** is a sequence of observations, $\{X_t\}$, one taken at each time t, and arranged in chronological order.



There are many methods available for modelling time series...

Missing Observations

...however, most methods require that the series is **contiguous** (no missing observations).

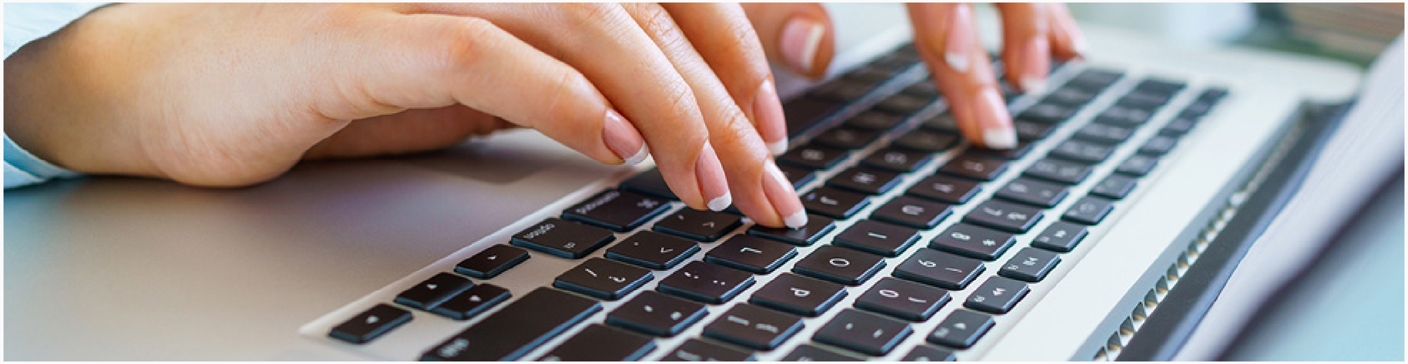


Thus, missing observations must be **imputed** (interpolated).

Why?

There are a number of reasons why a time series might have missing observations.

- Weekends or holidays
- Equipment failure
- Environmental constraints
- Transcription errors or incorrect data



Goals of the Project

Project completed for an undergraduate course, MATH4800H

- Research a variety of time series interpolation algorithms
- Evaluate their performances on a number of real-world datasets
- Use multiple performance criteria

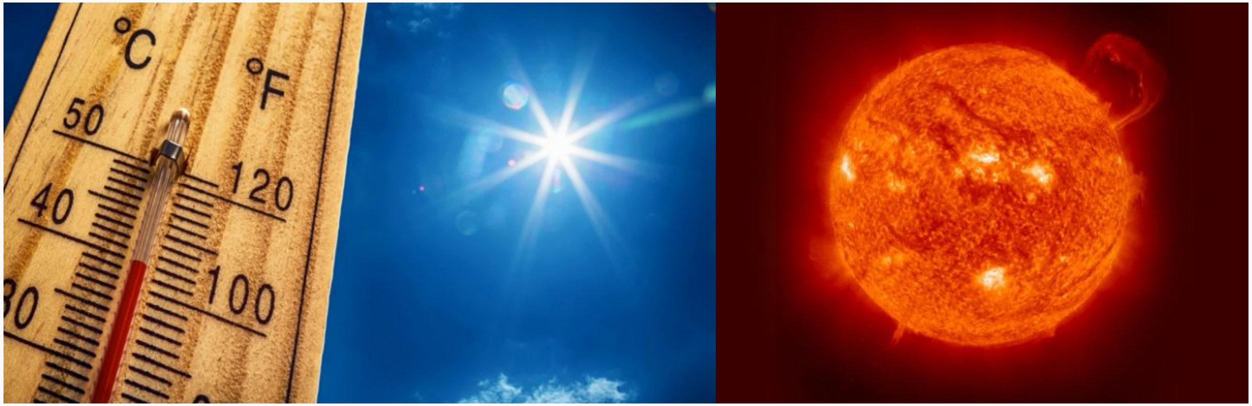
```
head(algorithm_names, n = 12)
```

```
##  [1] "Nearest Neighbor"
##  [2] "Linear Interpolation"
##  [3] "Natural Cubic Spline"
##  [4] "FMM Cubic Spline"
##  [5] "Hermite Cubic Spline"
##  [6] "Stineman Interpolation"
##  [7] "Kalman - ARIMA"
##  [8] "Kalman - StructTS"
##  [9] "Last Observation Carried Forward"
## [10] "Next Observation Carried Backward"
## [11] "Simple Moving Average"
## [12] "Linear Weighted Moving Average"
```

Datasets Used

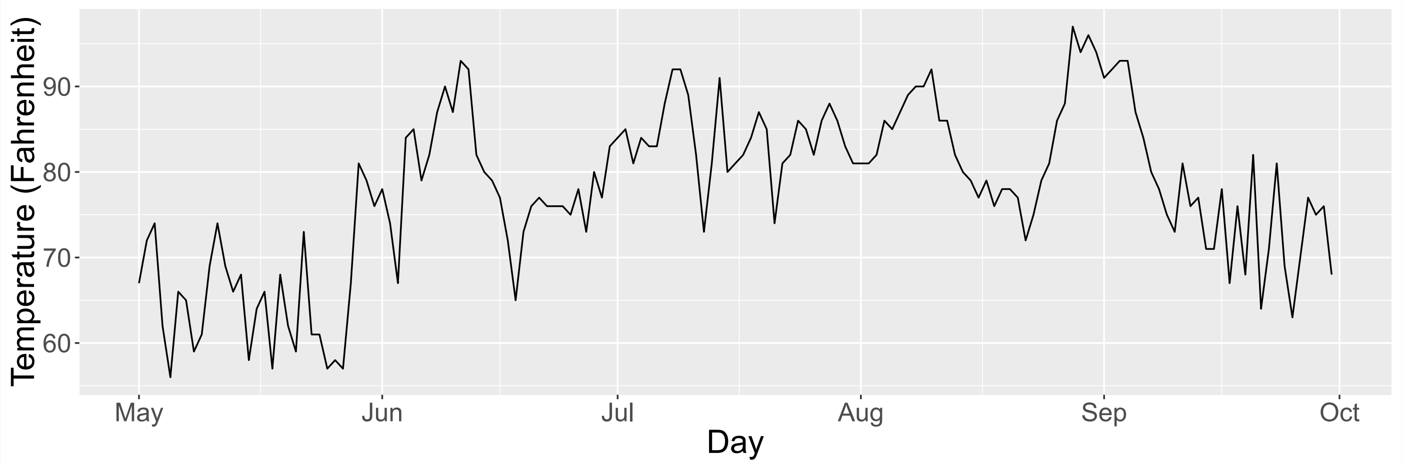
What kind of data were used?

- All time series used were real-world datasets (**not** simulated data)
- Non-stationary series
- Desirable to use series of varying length and spacing between observations
- Some of these will likely be familiar to some of you



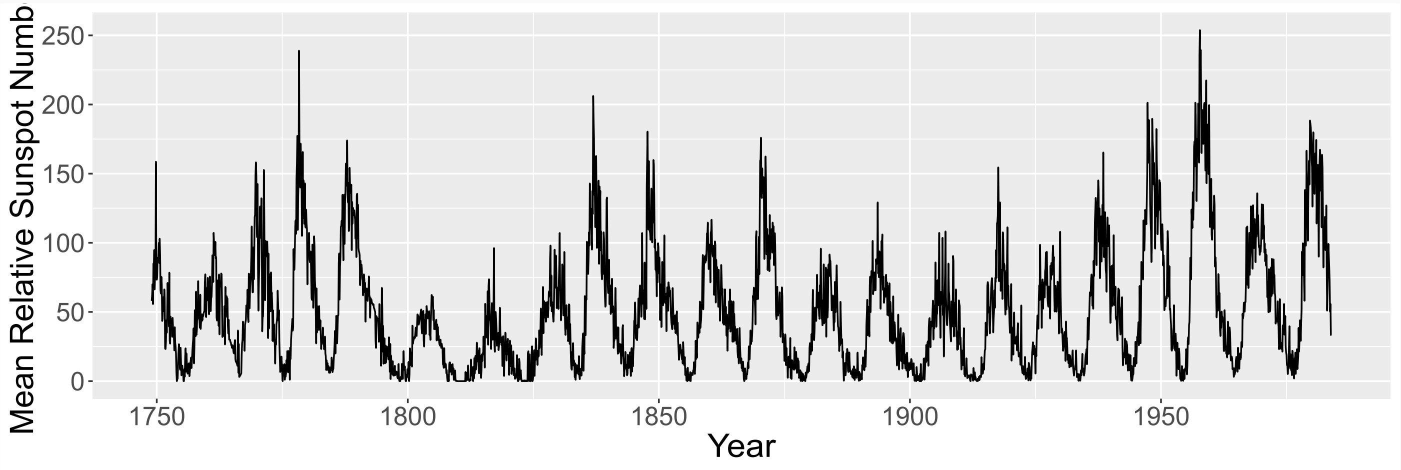
First Dataset

- Daily measurements of temperature (in Fahrenheit) in New York, May to September of 1973
- (The `temperature` variable from the `airquality` dataset in `R`)



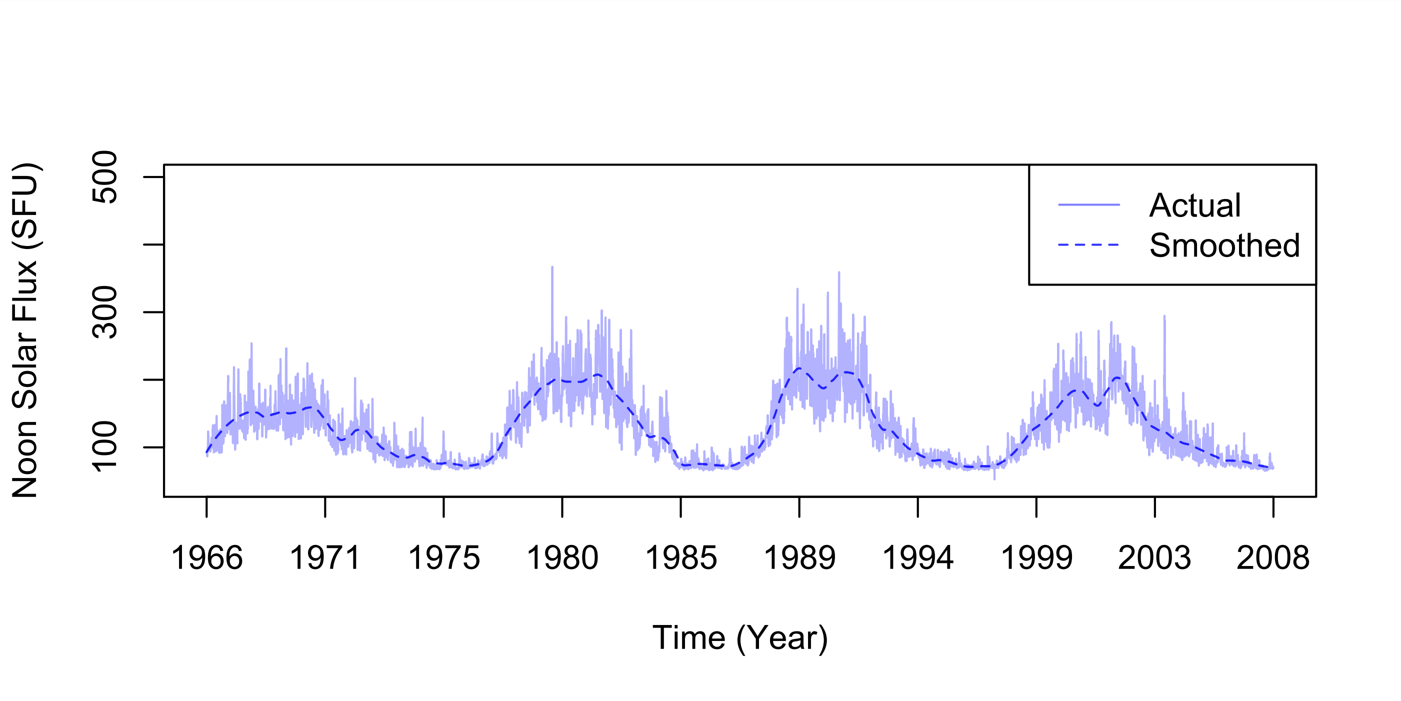
Second Dataset

- Monthly mean relative sunspot numbers from 1749 to 1983
- (The `sunspots` dataset in `R`)



Third Dataset

- Daily noon solar flux measurements from Penticton, British Columbia
- (The `PentOrig` variable from the `flux` dataset in the `tsinterp` package in `R`)



The Experiment

What was done?

Wrote an `R` script to do the following:

- Randomly impose gaps on each of the datasets (5%, 10%, 15%, 20%, 25%)
- Use 18 different interpolation algorithms to fill in the gaps
    - Including algorithms from the `R` packages `zoo`, `forecast`, `imputeTS`, and `tsinterp`
- Evaluate the performance of each algorithm using 17 different performance criteria
- Create tables summarizing the algorithms which performed the best and the worst (for each gap level, dataset, and performance criteria)

Results

- There wasn't one overall "winner", but in most cases, the best performing algorithms were:
    - Exponential Weighted Moving Averages
    - Kalman Filters
    - Cubic Splines
- The algorithms which performed well performed **very** well, and very comparably. For example: a subset of the 20% gap level results for r:

algorithm	airquality	sunspots	flux
Natural Cubic Spline	**0.98159651**	**0.98503122**	0.99901174
FMM Cubic Spline	0.98039913	0.98503142	0.99901174
Hermite Cubic Spline	0.97911815	0.9880365	0.99921698
Kalman - ARIMA	0.97012021	**0.99059037**	**0.99927336**
Kalman - StructTS	0.97049236	0.99048929	0.99927209
Linear Weighted MA	0.96826562	0.99011662	**0.99834509**
Exponential Weighted MA	0.97064295	0.99033871	0.99878023
Hybrid Wiener Interpolator	**0.96461487**	0.98974162	0.99856881

Next Steps

- Expand analysis to include more datasets
  - Will require a significant increase in computational power
- Experiment with varying gap lengths and gap selection methods
- Include datasets from a wide variety of fields
- Create recommendations for which algorithm to use based on the type of data

Want to learn more? Check out Sophie Castel's presentation on Monday!

14:15-14:30 **Sophie Castel** (Trent University), **Melissa Van Bussel** (Trent University), **Wesley Burr** (Trent University)
Imputation of Missing Values in Time Series Data / Imputation de valeurs manquantes dans des séries chronologiques

References

1. Mathieu Lepot, Jean-Baptiste Aubin, and Francois H.L.R. Clemens. Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance and Uncertainty Assessment. Water 2017, 9(10), 796.

2. Wesley S. Burr. Air Pollution and Health: Time Series Tools and Analysis. Queen's University, PhD thesis. 2012.

3. Wesley S. Burr (2012). `tsinterp`: A Time Series Interpolation Package for `R`. R Package.

Thank You

melissavanbussel@trentu.ca

linkedin.com/in/melissavanbussel

github.com/melissavanbussel

Slides created via the R package xaringan. Slides and accompanying files are available on GitHub.