Imputing average temperatur

Maxime Rischard[*]

[*]Department of Statistics, Harvard University

February 24, 2018

Executive summary

- In 2015, the average temperature at Waterloo Municipal Airport was 9.497 ° C.
- Extracting daily $\mathrm{T}_n/\mathrm{T}_x$ at 17:00 UTC each day and taking their mean yields an estimate of the average temperature of 9.716 ° C.
- Taking the mean of the imputed temperatures yields a posterior distribution of the average temperature of 9.474 ° C ± 0.020.
- That's pretty good!

Background

In the paper presenting the temperature imputations approach to the hour of measurement bias, we demonstrated our ability to "recover" summary statistics such as the average maximum temperature. We focused on statistics such as the average daily maximum temperature, that are deterministic functions of the daily extrema, which are in turn functions of the true temperatures and hour of measurement. To formalize this a little bit, let $\mathrm{T}_{\mathrm{miss}}$ be the time series of temperatures at the location of interest, and `hour` be the hour of measurement. Then we can write the daily temperature extrema as functions $\mathrm{T}_n\left(\mathrm{T}_{\mathrm{miss}}, \mathtt{hour}\right)$ and $\mathrm{T}_x\left(\mathrm{T}_{\mathrm{miss}}, \mathtt{hour}\right)$. We additionally used information from temperature time series at nearby airport, which we denote $\mathrm{T}_{\mathrm{nearby}}$. The paper shows how to draw imputations of $\mathrm{T}_{\mathrm{miss}}$ from its posterior distribution $\mathrm{T}_{\mathrm{miss}} \mid \mathrm{T}_n, \mathrm{T}_x, \mathrm{T}_{\mathrm{nearby}}, \mathtt{hour}$. The summary statistics of interest are of the form $h\left(\mathrm{T}_n, \mathrm{T}_x\right)$, which we can expand to $h\left(\mathrm{T}_n\left(\mathrm{T}_{\mathrm{miss}}, \mathtt{hour}\right), \mathrm{T}_x\left(\mathrm{T}_{\mathrm{miss}}, \mathtt{hour}\right)\right)$. Therefore the value of the summary statistic indirectly depends on the hour of measurement. In the paper, we show we can infer reasonably well what the statistic *would have been*, had we made the measurements at a different time $\mathtt{hour}'$. That is, we successfully obtain samples from the posterior distribution

$$h\left(T_n\left(T_{\text{miss}},\texttt{hour}'\right),T_x\left(T_{\text{miss}},\texttt{hour}'\right)\right)\mid T_n\left(T_{\text{miss}},\texttt{hour}\right),T_x\left(T_{\text{miss}},\texttt{hour}\right),T_{\text{nearby}},\texttt{hour} \tag{1}$$

simply by applying the function $h\left(T_n\left(T_{\text{miss}},\texttt{hour}'\right),T_x\left(T_{\text{miss}},\texttt{hour}'\right)\right)$ to the imputations of T_{miss} we had previously obtained.

New Problem

In *this* document, we are interested in the year's average temperature, which is a slightly different kind of summary statistic. Instead of being of function of T_n and T_x, it is a function directly of T_{miss}, and does not depend on `hour`, so we can write it simply as $m\left(T_{\text{miss}}\right)$. However, since T_{miss} is not available, the average temperature is commonly *estimated approximately* by a statistic of the first form, the mean of the minima and maxima:

$$m\left(T_{\text{miss}}\right)\approx h\left(T_n,T_x\right)=\left(\overline{T_n}+\overline{T_x}\right)/2\,. \tag{2}$$

The estimate depends on `hour` but the estimand does not. What we propose is simply to once again obtain the posterior distribution of $m\left(T_{\text{miss}}\right)$ by applying the m function to the imputations, which gives us samples from

$$m\left(T_{\text{miss}}\right)\mid T_n\left(T_{\text{miss}},\texttt{hour}\right),T_x\left(T_{\text{miss}},\texttt{hour}\right),T_{\text{nearby}},\texttt{hour} \tag{3}$$

True mean temperature at Waterloo Airport

Because I already have imputations, it was easiest to apply this idea to the Waterloo Municipal Airport. In that case I actually have access to hourly measurements, so I can compute the true $m\left(T_{\text{miss}}\right)$. The time series is provided as $N=12{,}695$ measurements $T_{\text{miss},i}$, $i=1,\ldots,N$ made at times t_i. I refer to these as hourly measurements, but in reality the time series is irregular, with median difference of 54 minutes, mean difference 41.4 minutes, minimum difference 1 minute, and maximum difference 120 minutes (histogram in Figure 1). For this reason, we can't simply average the temperature measurements, we should weight each observation by the amount of time it occupies. This gives weight

$$w_i=\begin{cases}\frac{1}{2}\left(t_{i+1}-t_i\right) & ,i=1\\ \frac{1}{2}\left(t_{i+1}-t_{i-1}\right) & ,i=2,\ldots,N-1\\ \frac{1}{2}\left(t_i-t_{i-1}\right) & ,i=N\end{cases} \tag{4}$$

to the i^{th} observation $T_{\text{miss},i}$. Another way to write this, which turns out to be notationally more conve-

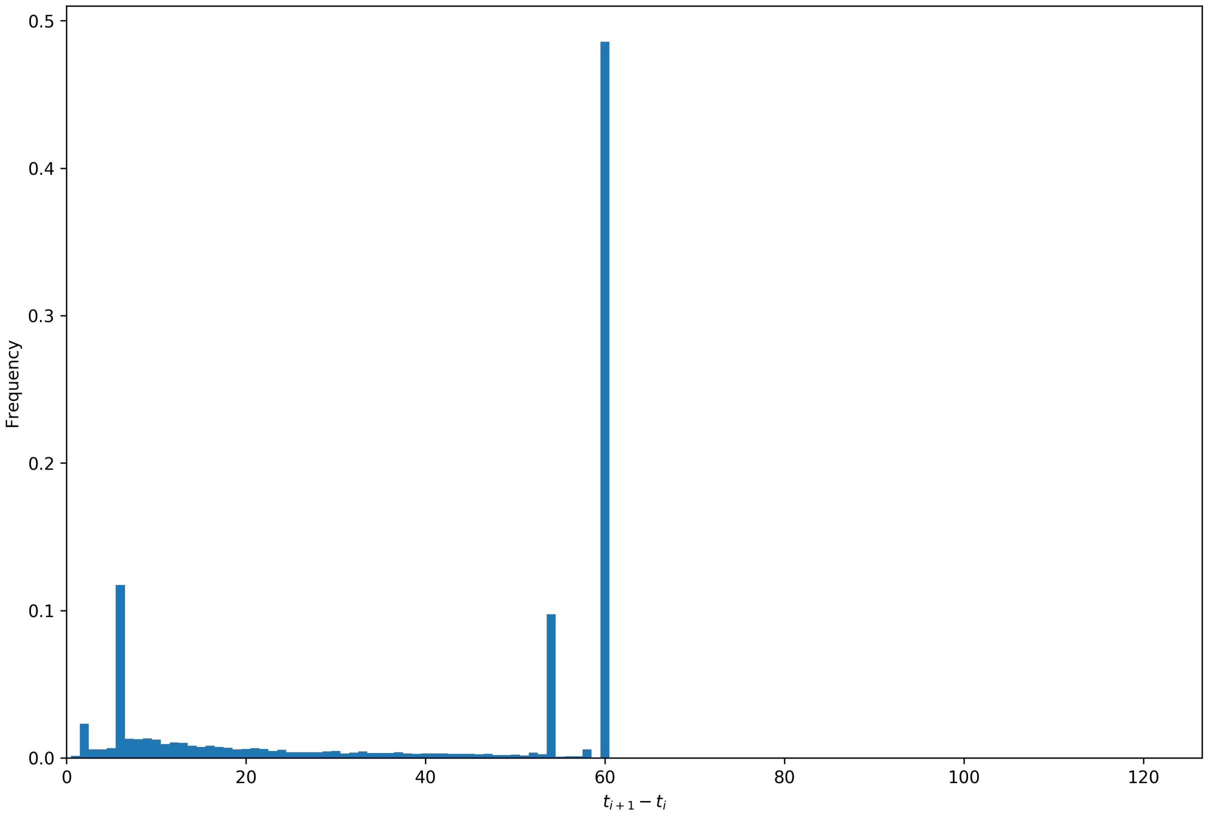

Figure 1: *Histogram of time differences*

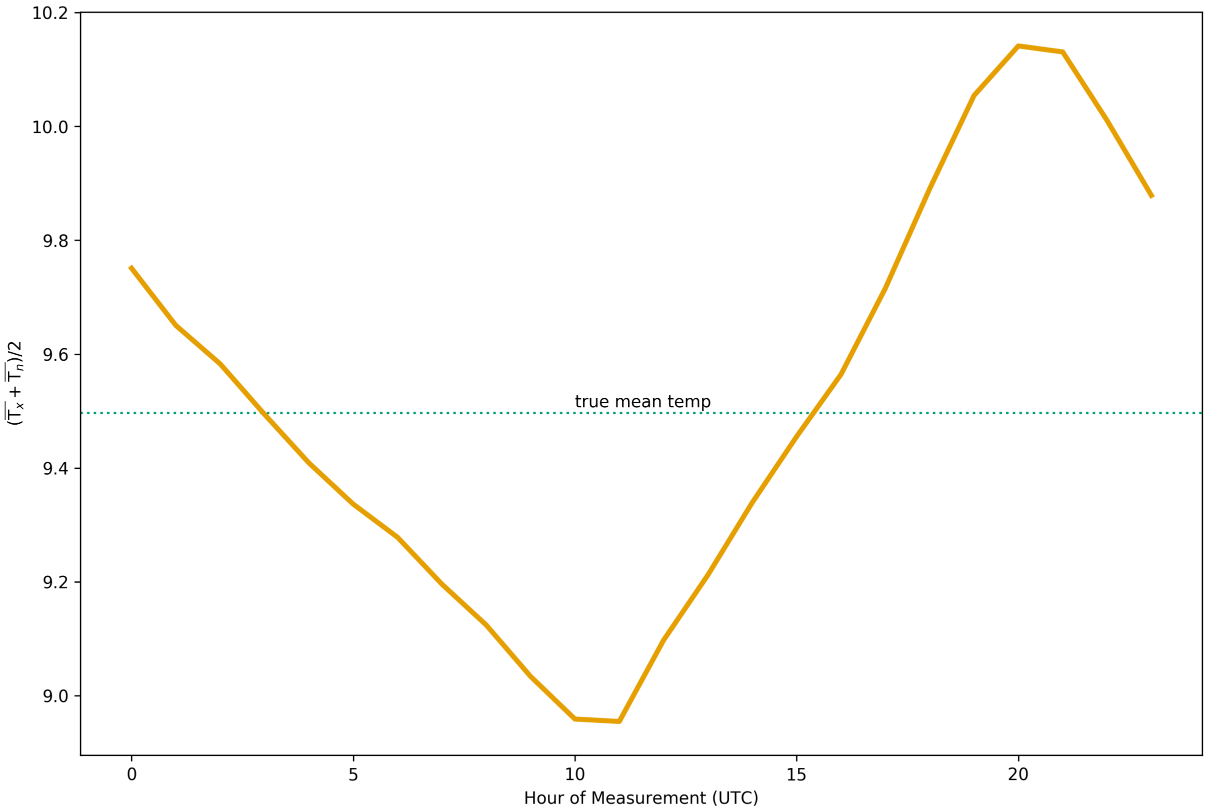

Figure 2: *The 2015 mean temperature at Waterloo Municipal Airport conventionally estimated by averaging* T_n *and* T_x *depends on the hour of measurement.*

nient, is to weight the temperature midpoints

$$\text{middle}\left(T_{\text{miss}}, i\right) \equiv \left(T_{\text{miss},i+1} + T_{\text{miss},i}\right)/2\,, i = 1, \ldots, N-1 \tag{5}$$

by the time differences

$$\Delta_i = t_{i+1} - t_i\,. \tag{6}$$

The weighted mean can therefore be written as:

$$m\left(T_{\text{miss}}\right) \approx m\left(\{\left(T_{\text{miss},i}, t_i\right) : i = 1, \ldots, N\}\right) \equiv \frac{1}{t_N - t_1} \sum_{i=1}^{N-1} \text{middle}\left(T_{\text{miss}}, i\right) \Delta_i \tag{7}$$

Applying this weighted mean formula to the temperatures measured at Waterloo Municipal Airport yields an annual mean of $m\left(T_{\text{miss}}\right) = 9.497\,^{\circ}$ C in 2015.

Conventional estimate of mean temperature at Waterloo Airport

The conventional approach to estimating $m\left(T_{\text{miss}}\right)$ is to average the daily T_n and T_x measurements. At Waterloo Airport, since we have T_{miss}, we can emulate this by first obtaining $T_n\left(T_{miss}, \texttt{hour}\right)$, $T_x\left(T_{miss}, \texttt{hour}\right)$ and then averaging the results together. The result depends on `hour`, and so I show the results as a function of `hour` in Figure 2. Strikingly, the estimate varies by over 1 ° C depending on the hour of measurement. If the T_n/T_x measurements are made at 17:00 UTC, the average temperature will be estimated as 9.716 ° C.

Imputation-based estimate of mean temperature at Waterloo Airport

Now I go ahead and apply the m function to the imputations. This isn't quite so straightforward because for computational reasons I obtained the temperature imputations for 2015 in nine-day windows, with three days overlap between adjacent windows. What follows is some mildly painful notation to make it completely clear how I'm obtaining my imputation-based estimate. We denote the time interval covered by the kth window as $\texttt{win}_k$, so $\{i : t_i \in \texttt{win}_k\}$ is the indices of the observations that are within this window. Each observation can fall in multiple windows. For a given time t and a window k, I can calculate the buffer $\texttt{buff}(t;\texttt{win}_k) = \min\left(\max\left(\texttt{win}_k\right) - t, t - \min\left(\texttt{win}_k\right)\right)$ that separates the time t and the window's closest edge. The best window for imputing an observation at time t is then found by $\arg\max_k\left(\texttt{buff}\left(t;\texttt{win}_k\right)\right)$. And the set of indices of the midpoints (times halfway between two observations) for which window k is the best is then written as:

$$\texttt{best}_k \equiv \left\{ i : \arg\max_{k'}\left(\texttt{buff}\left(\mathrm{middle}\left(t, i\right); \texttt{win}_{k'}\right)\right) = k \right\} . \tag{8}$$

Those midpoints occupy a duration of $\Delta\left(\texttt{best}_k\right) = t_{\max\left(\texttt{best}_k\right)+1} - t_{\min\left(\texttt{best}_k\right)}$. For the l^{th} imputed time series $T_{miss}^{(l)}$ $(l = 1, \ldots, L)$ within each window k, I obtain the weighted mean of imputed temperatures for the midpoints in the middle of the window:

$$\overline{T}_k^{(l)} = m\left(\left\{\left(T_{miss,i}^{(l)}, t_i\right) ; i \in \texttt{best}_k\right\}\right) , \tag{9}$$

as defined in (7). The posterior mean and variance μ_k, σ_k^2 are estimated by computing the sample mean and variance of $\overline{T}_k^{(l)}$ over the L imputations. Then I combine the estimates for each window by taking a weighted mean of the μ_k's:

$$\begin{aligned} \mathbb{E}\left(m\left(T_{miss}\right) \mid T_n, T_x, T_{nearby}, \texttt{hour}\right) &= \frac{1}{t_N - t_1} \sum_{k=1}^{K} \mu_k \, \Delta\left(\texttt{best}_k\right) \\ \mathrm{var}\left(m\left(T_{miss}\right) \mid T_n, T_x, T_{nearby}, \texttt{hour}\right) &\approx \frac{1}{\left(t_N - t_1\right)^2} \sum_{k=1}^{K} \sigma_k^2 \, \Delta\left(\texttt{best}_k\right)^2 \end{aligned} \tag{10}$$

The expectation of the posterior of $\mathfrak{m}\left(T_{\text{miss}}\right)$ is exact up to Monte Carlo error, but the variance estimate ignores correlations between windows. Long story short, the resulting estimate is

$$\begin{aligned} \mathbb{E}\left(\mathfrak{m}\left(T_{\text{miss}}\right) \mid T_{n}, T_{x}, T_{\text{nearby}}, \texttt{hour}\right) &= 9.474^{\circ}\,\text{C} \\ \text{var}\left(\mathfrak{m}\left(T_{\text{miss}}\right) \mid T_{n}, T_{x}, T_{\text{nearby}}, \texttt{hour}\right) &\approx \left(0.020^{\circ}\,\text{C}\right)^{2}\,, \end{aligned} \tag{11}$$

which is a good estimate of the true $\mathfrak{m}\left(T_{\text{miss}}\right) = 9.497^{\circ}$ C.

References