TemperatureImputations

Maxime Rischard

December 17, 2016

Contents

1 Introduction

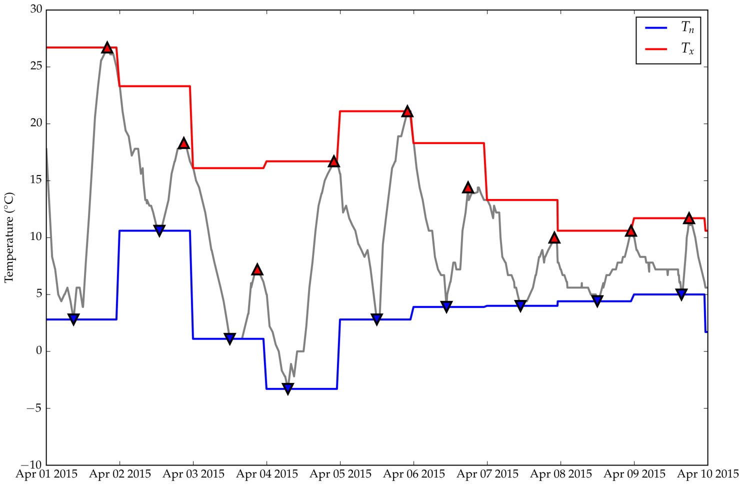
- explain the problem we're trying to solve

1.1 bias in recorded daily minima and maxima induced by time of measurement

- explain
- demonstrate using hourly temperatures from one station: reduce to daily min and max and show difference as a function of measurement hour

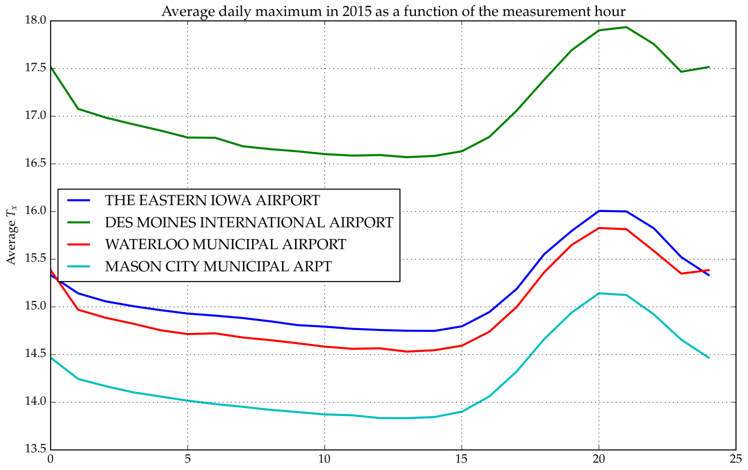
We illustrate the measurement bias in daily maxima and minima with ten days of hourly temperature measurements from the Waterloo Municipal Airport station in Iowa. Ideally, T_x measurements should capture the peak of each diurnal cycle, and T_n its trough. In Figure X, those ideal measurements are indicated by the red and blue triangles respectively. The actual measurements are obtained by dividing the data into 24 hour measurement windows, and extracting the minimum and maximum. For each window, we plot these extrema with a red and blue horizontal line.

Most of the time, the ideal measurement and the actual measurement coincide: the triangle is on that day's line. But there are also several misses. The most blatant example occurs on April 3rd, where the peak of the diurnal cycle is {{apr3_realmax}}°C and occurs at 21:00 UTC. However, because the previous day was much warmer, the day's T_x record of {{apr3_measured}}°C is reached immediately after the previous day's measurement. The measured T_x therefore overestimates the diurnal cycle's peak by {{round(apr3_measured-apr3_realmax,1)}}°C.

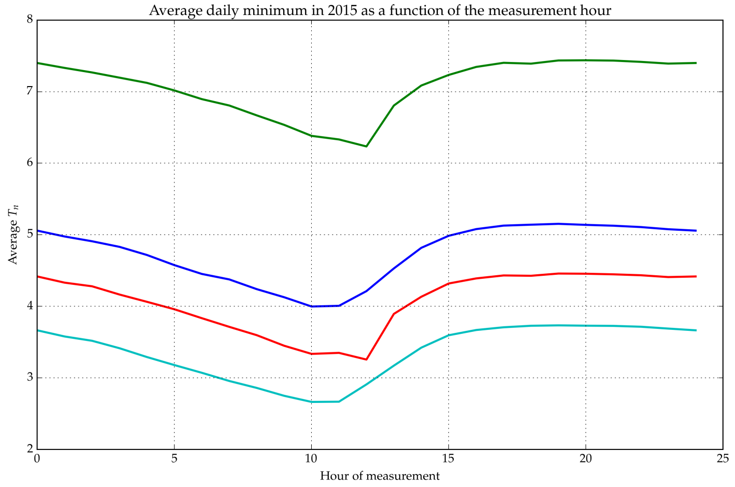


This subtle bias in the daily records can in turn bias long-term summary statistics that are of climatological interest. A measure as simple as the average daily maximum temperature for an entire year (2015) increases by over 1°C if the measurements are made at the warmest time of day 21:00 UTC rather than 14:00 UTC (see Figure X). Conversely, the average T_n is colder by over 1°C if T_n is measured at 10:00 UTC (the coldest time of day on average) rather than 17:00 UTC.

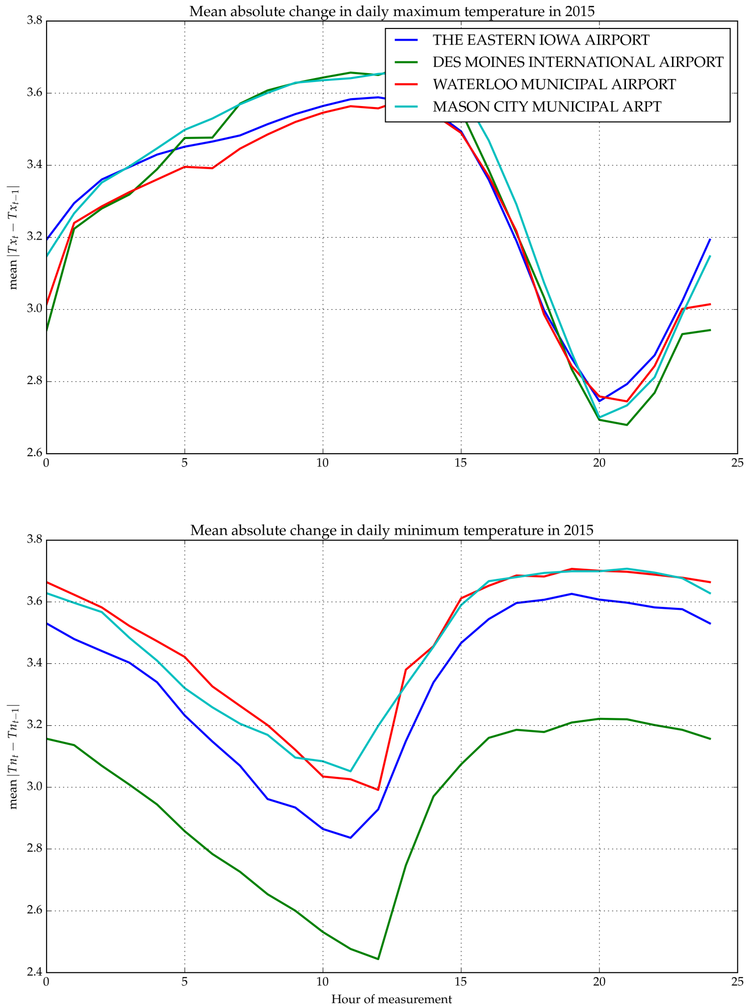
A climatologist studying weather variability might be interested in summary statistics such as the average absolute change in the daily temperature maxima and minima from one day to the next. The answer to that question too depends on the time of day at which the temperatures are recorded. Collecting the measurements at the hottest time of day means that the peaks on a warm day gets recorded twice, erasing the diurnal peaks of the following colder day, and hence the variability gets underestimated. We can see this in Figure X, where the respective variability estimates drop if the maxima get measured at the warmest time, or if the minima get measured at the coldest time.



Average daily maximum in 2015 as a function of the measurement hour



Average daily minimum in 2015 as a function of the measurement hour



1.2 Proposed solution

We have seen that the daily maxima and minima do not faithfully record each diurnal cycle's peak and trough. The peaks on a relatively cold day can get overwritten by temperatures at either end of the measurement window that properly belong to the previous or the next diurnal cycle. Troughs on relatively warm days can be similarly overwritten. Our goal is to undo this damage, and recover estimates of summary statistics, such as the average daily maximum temperature, that do not suffer from the consequent bias. We need to address the erasure of information caused by the measurement mechanism, and therefore view this as a missing data problem.

Taking the missing data perspective, we seek to impute the hourly temperatures that have been replaced by a maximum and minimum over a 24 hour period. To do so, we use information from two sources: the recorded daily temperature extremes at the station of interest, and also hourly temperatures recorded at nearby meteorological stations. These hourly measurements are considered less reliable by climatologists, as they aren't as carefully documented, calibrated, and situated. The meterological stations are often in locations (like airports) where human activity will affect temperatures. Therefore, summary statistics extracted directly from those measurements would not be directly usable for climatology, as they could suffer from systematic bias. However, even if miscalibrated, the meterological data do contain valuable information about the hourly changes in temperatures on any given day. We therefore use them to inform the shape of the imputed temperature time-series at our location of interest, while we use the recorded temperature extrema to calibrate and constrain them.

2 First Spatiotemporal Model

To model measured temperatures at various locations and times, we use a spatio-temporal Gaussian process model. In its simplest form, we believe that temperatures from stations that are near each other are more correlated than distant stations, and that those correlations should also decay in time. In the spatial statistics literature, squared exponential covariance functions are commonly used to model correlations decaying as a function of distance. Ignoring the time dimension, we would model the simultaneous temperatures throughout a region as a Gaussian process, with the covariance of two locations $\mathbf{x}$ and $\mathbf{x}'$

$$\operatorname{cov}\left(T(\mathbf{x}), T(\mathbf{x}') \mid t\right) = k_{space}(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{GP}}^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^{\intercal}(\mathbf{x}-\mathbf{x}')}{2\ell_x^2}\right). \tag{1}$$

Similarly, ignoring the spatial dimension, the time series of temperatures at a single location can be modeled as a Gaussian process with covariance between two moments t and t'

$$\operatorname{cov}\left(T(t), T(t') \mid \mathbf{x}\right) = k_{time}(t, t') = \sigma_{\mathrm{GP}}^2 \exp\left(-\frac{(t-t')^2}{2\ell_t^2}\right). \tag{2}$$

We then combine the spatial and temporal model by multiplying the covariance functions

$$k_{st}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}'). \tag{3}$$

This gives us the covariance of the Gaussian process underlying the full spatio-temporal model of temperatures. To complete the model specification, we add a mean temperature for each station $\mu_{\mathrm{station}[i]}$, and iid measurement noise ϵ_i.

$$\begin{aligned} T_i &= \mu_{\mathrm{station}[i]} + f(\mathbf{x}_i, t_i) + \epsilon_i && (4)\\ f(\mathbf{x}_i, t_i) &\sim \mathcal{GP}\left(0, k_{st}(\mathbf{x}, \mathbf{x}', t, t')\right) && (5)\\ \epsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right) && (6)\\ & && (7) \end{aligned}$$

3 Fitting the spatiotemporal model

Software is readily available in many programming languages for fitting Gaussian process models, including inference on the covariance parameters. We chose to use the julia `GaussianProcesses.jl` package to fit the above spatiotemporal model to the hourly temperatures at four Iowa weather stations. However, the Iowa data includes X measurements, which is computationally infeasible to fit directly with a single Gaussian process. While approximation techniques exist to fit such large datasets, we chose the less efficient but simpler approach of dividing the data into 10-day chunks, modeled as independent Gaussian processes with shared hyperparameters. To simplify our implementation, we replaced the $\mu_{\mathrm{station}}\,[i]$ terms by a spatial squared exponential component

$$k_{\mu}(\mathbf{x}, \mathbf{x}') = \sigma_{\mu}^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^{\intercal}(\mathbf{x}-\mathbf{x}')}{2\ell_{\mu}^2}\right) \tag{8}$$

with large variance σ_{μ}^2 and low lengthscale ℓ_{μ} added to the covariance function so that the spatio-temporal kernel becomes

$$k_{st}(\mathbf{x}, \mathbf{x}', t, t') = k_{time}(t, t') \cdot k_{space}(\mathbf{x}, \mathbf{x}') + k_{\mu}(\mathbf{x}, \mathbf{x}'). \tag{9}$$

The entire model therefore has 3 parameters: σ_{GP}, ℓ, and σ_ϵ. By optimizing the marginal likelihood of the Iowa data as a function of these three parameters, we obtained $\sigma_{\mathrm{GP}} = 3.73\,^{\circ}\mathrm{C}$, $\ell = 176.4\,\mathrm{km}$ and $\sigma_\epsilon = 0.44\,^{\circ}\mathrm{C}$.

1. timeseries model
   - fitting hyperparameters
   - chunks
   - show variogram
2. spatiotemporal model
   - fitting hyperparameters
   - chunks
3. imputations
   - Stan
   - softmin and softmax
   - observation noise
   - reparametrization

4 Improving model

1. focused on timeseries model
   - kernel components
   - diurnal cycle
   - show improved variograms
2. spatiotemporal model
   - variograms and cross-variograms
   - trace evolution
     - product kernel
     - sum of products with variance 1
     - sum of products with free variance
   - for each model, report marginal likelihood, and predictive diagnostic in a table
   - discuss importance of getting uncertainty right

5 Analysis

- show imputations on interesting days
- show imputations can capture two possible explanations for a measurement
- discuss possibility of inferring measurement time