TemperatureImputations

Maxime Rischard

February 24, 2017

Contents

1 Introduction

- explain the problem we're trying to solve

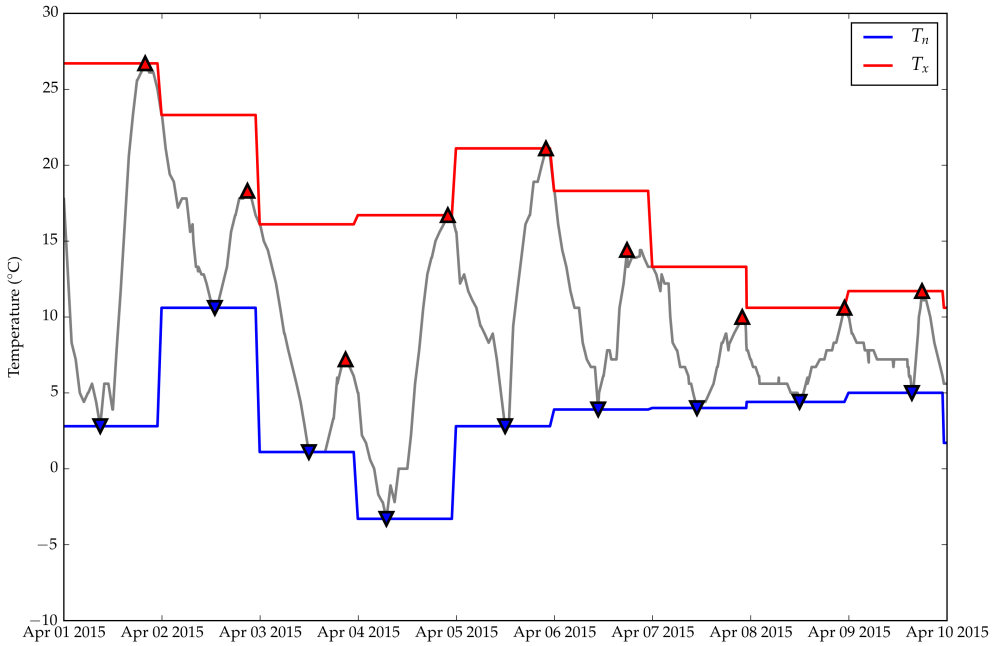
1.1 bias in recorded daily minima and maxima induced by time of measurement

- 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.

On most days, 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.



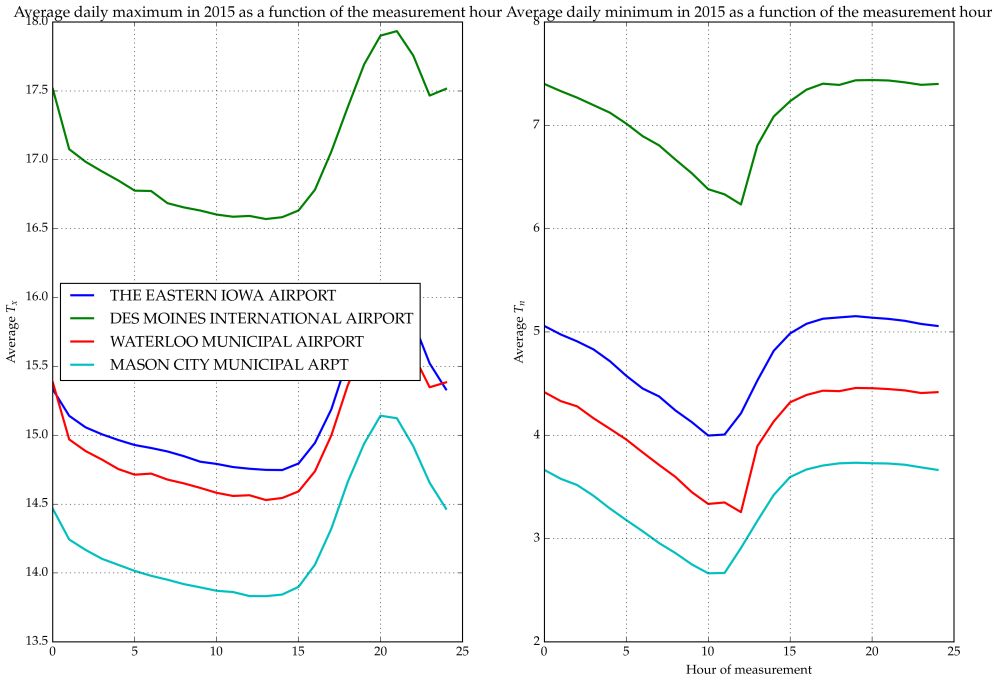
This subtle bias in the daily records can in turn bias long-term summary statistics that are of climatological interest. A measure as simple as the average daily maximum temperature for an entire year (2015) increases by over 1°C if the measurements are made at the warmest time of day 21:00 UTC rather 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.

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

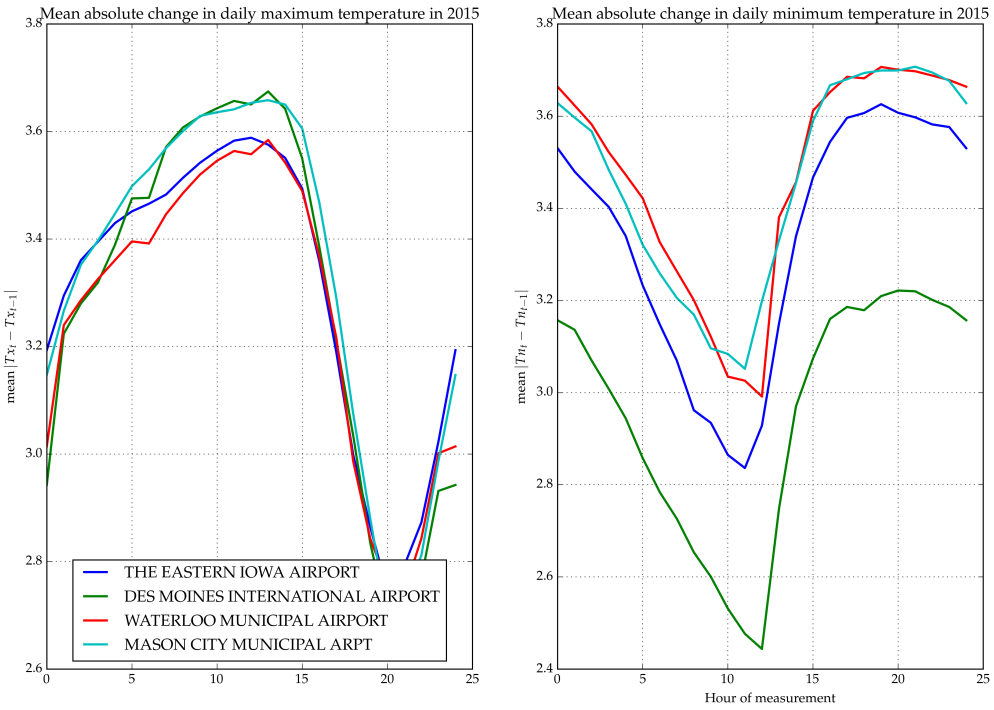
To model measured temperatures at various locations and times, we use a spatio-temporal Gaussian process model. In its simplest form, we believe that temperatures from stations that are near each other are more correlated than distant stations, and that those correlations should also decay in time. In 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 $\mathbf{x}'$

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

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'

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

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 $\mu_{\text{station}[i]}$, and iid measurement noise ϵ_i.

$$T_i = \mu_{\text{station}[i]} + f(\mathbf{x}_i, t_i) + \epsilon_i \tag{4}$$

$$f(\mathbf{x}_i, t_i) \sim \mathcal{GP}\left(0, k_{st}(\mathbf{x}, \mathbf{x}', t, t')\right) \tag{5}$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right) \tag{6}$$

$$\tag{7}$$

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 47,864 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 $\mu_{\text{station}}[i]$ 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

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

The model therefore has 4 free parameters: σ_{GP}, ℓ_t, ℓ_x and σ_ϵ. By optimizing the marginal likelihood of the Iowa data as a function of these three parameters, we obtained $\sigma_{GP} = 3.73\,°C$, $\ell_t = 2.7\,\text{hours}$, $\ell_x = 176.4\,\text{km}$ and $\sigma_\epsilon = 0.44\,°C$.

1. timeseries model
    - fitting hyperparameters
    - chunks
    - show variogram
2. spatiotemporal model
    - fitting hyperparameters
    - chunks

4 Imputations

```
* Stan
* softmin and softmax
* observation noise
* reparametrization
```

Once we have a spatio-temporal Gaussian process model with optimized covariance parameters, we can use it to generate predictions at the station where we aim to generate imputations based on nearby measurements. Gaussian processes make this a closed-form procedure. We'll denote the temperatures we wish to impute as T_{miss} at times t_{miss} and location $\boldsymbol{x}_{miss}$ and those observed at nearby stations as T_{obs}, at times t_{obs} and locations X_{obs}. Under the spatio-temporal model, T_{miss} and T_{obs} are jointly multivariate normal, with mean zero and covariance given by $k_{st}(\boldsymbol{x}, \boldsymbol{x}', t, t')$. Standard results for conditioning within multivariate normals then yields

$$T_{miss} \mid T_{obs} \sim \mathcal{N}\left(\mu_{miss|obs}, \Sigma_{miss|obs}\right) \,, \tag{10}$$

$$\mu_{miss|obs} = \mathbb{E}\left(T_{miss} \mid T_{obs}\right) = \text{cov}\left(T_{miss}, T_{obs}\right) \text{cov}\left(T_{obs}, T_{obs}\right)^{-1} T_{obs} \,, \tag{11}$$

$$\Sigma_{miss|obs} = \text{var}\left(T_{miss} \mid T_{obs}\right) = \text{cov}\left(T_{miss}, T_{miss}\right) - \text{cov}\left(T_{miss}, T_{obs}\right) \text{cov}\left(T_{obs}, T_{obs}\right)^{-1} \text{cov}\left(T_{obs}, T_{miss}\right) \,. \tag{12}$$

$$\tag{13}$$

All covariance matrices can be obtained by plugging into k_{st}. For example, the ijth entry of $\text{cov}\left(T_{miss}, T_{obs}\right)$ is given by $k_{st}(\boldsymbol{x}_{miss}, X_{obs}[j], t_{miss}[i], t_{obs}[j])$, where $X_{obs}[j]$ gives the spatial covariates of the jth observation, and $t_{obs}[j]$ its time.

In Figure X, we show an example of predictions obtained from this spatio-temporal model. We witheld measurements from the Waterloo Municipal Airport, and then used data from three nearby stations between May 25, 2015 and June 3, 2015 to predict the Waterloo temperatures during the same time window. This setup allows us to gauge the quality of the predictions.

Our aim, however, isn't just to predict temperatures at a location with no measurements, but rather to impute hourly temperatures at a location with accurate measurements of the daily temperature extremes. Those measurements can be thought of as constraints on the predictions. A T_x record means that in the 24 hours before the measurements, the temperature reached T_x but never exceeded it. Conceptually, we could therefore implement a valid imputation algorithm by drawing random samples from the posterior predictive multivariate normal distribution obtained from nearby measurements, and only keeping the samples that satisfy this constraint. Unfortunately, the probability of a random draw exactly satisfying

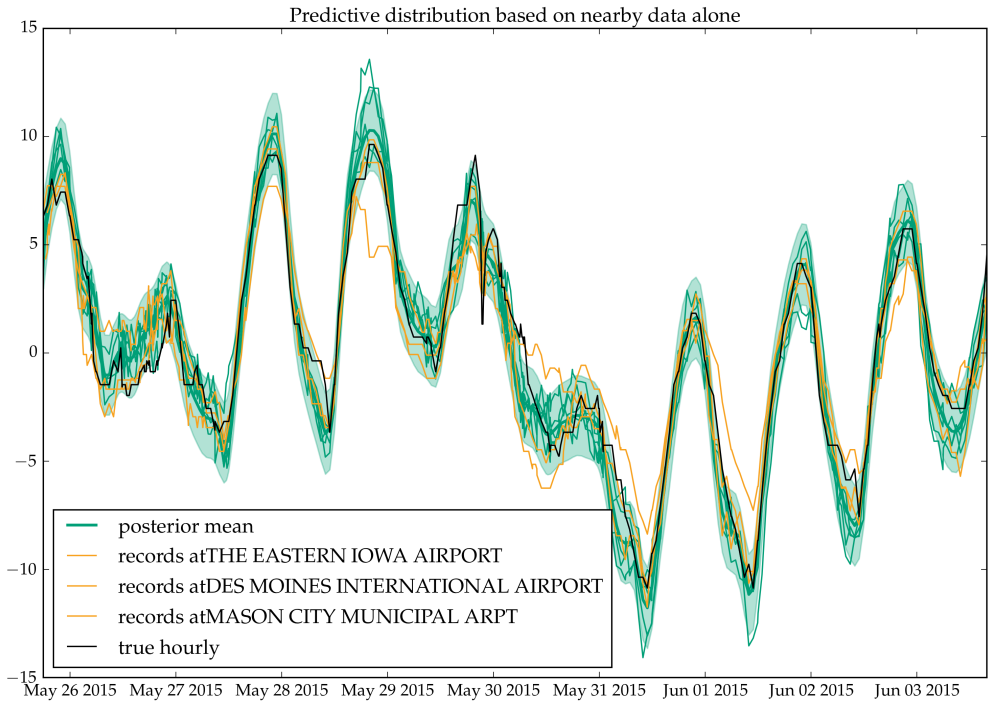


Figure 1: Predictive distribution using only nearby data and the simple product of square exponentials model. The orange lines are the measurements at nearby stations that are being used to inform the predictions. The black line is the true temperatures that have been withheld from the model, while the green line is the posterior mean of the predictions. The credible range (in green) is twice the standard deviations extracted from the diagonal entries of the posterior covariance matrix.

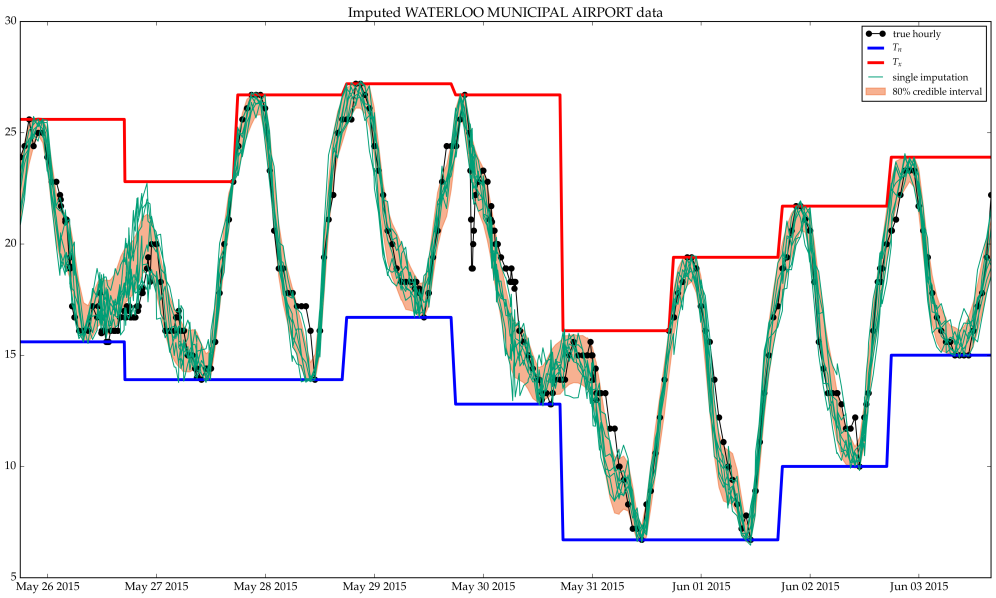


Figure 2: Imputations at Waterloo Airport from May 25, 2015 to June 3, 2015

such a constraint is zero, and so there would need to be some tolerance for overshooting or undershooting each day's T_n and T_x constraints. Imputing longer periods would then require either increasing the number of samples exponentially, or further loosening this tolerance. Ultimately, this rejection sampling strategy is therefore bound to fail.

Instead, we used the probabilistic programming language Stan to draw samples from the constrained predictive distribution. In Stan, we specify a probabilistic data-generating process for the observed temperatures, based on parameters and latent variables with accompanying priors. Stan then uses a Hamiltonian Monte Carlo (HMC) algorithm to draw sample from the posterior distribution for the parameters and latent variables. In our case, the observations are the daily maxima and minima, the only parameter is the average temperature at the station of interest μ_{miss}, and the latent variables are the missing unobserved hourly temperatures T_{miss}.

[Serious notation problems here:]

$$\mu_{miss} \sim \mathcal{N}(0, 100) \tag{14}$$
$$f_{miss} \sim \mathcal{N}\left(\mu_{miss|obs}, \Sigma_{miss|obs}\right) \tag{15}$$
$$T_{miss} \sim \mu_{miss} + f_{miss} \tag{16}$$
$$T_x[d] = \max\{T_{miss,i} : t_{miss,i} \in d\} \tag{17}$$
$$T_n[d] = \min\{T_{miss,i} : t_{miss,i} \in d\} \tag{18}$$
$$\tag{19}$$

The problem lies in the sharpness of the max and min functions. At each step of the markov chain, a proposal is made for T_{miss}. If the daily maxima and minima of this proposal coincide exactly with the measurements, then its posterior probability is finite. Otherwise it is exactly zero, with gradient also zero. HMC works by exploiting the gradient of the posterior, and therefore fails to converge in this situation, for reasons similar to the failure of the naive rejection sampler.

We can rescue the algorithm by replacing the max and min functions with the smoothmax and smoothmin functions, which take real inputs $x_1, \ldots, x_p$ and a sharpness parameter k and return

$$\text{smoothmax}\left(x_1, \ldots, x_p; k\right) = \log\left(\sum_{i=1}^{p} e^{k x_i}\right) \tag{20}$$

$$\text{smoothmin}\left(x_1, \ldots, x_p; k\right) = -\,\text{smoothmax}\left(-x_1, \ldots, -x_p; k\right) \tag{21}$$

As $k \to \infty$, smoothmax becomes the maximum, and smoothmin becomes the minimum. Lower values of k give close approximations. When smoothmax replaces max and smoothmin replaces min, there is a small price in precision due to the approximation, but there is a huge computational benefit: gradients are now available. Furthermore, by adding a small amount of white noise with variance σ_ϵ to the T_x and T_n measurements, we ensure that the posterior doesn't go abruptly to zero when the conditions aren't met exactly. These modifications make HMC a viable algorithm to efficiently draw samples from the posterior. Setting k and σ_ϵ is a compromise between exactness and efficiency; we found $k = 10$ and $\sigma_\epsilon = 0.1$ to perform well. The full model we implemented is below, and is an approximation of our ideal model above.

$$\mu_{\text{miss}} \sim \mathcal{N}\left(0, 100\right) \tag{22}$$

$$f_{\text{miss}} \sim \mathcal{N}\left(\mu_{\text{miss|obs}}, \Sigma_{\text{miss|obs}}\right) \tag{23}$$

$$T_{\text{miss}} \sim \mu_{\text{miss}} + f_{\text{miss}} \tag{24}$$

$$T_x\left[d\right] \sim \mathcal{N}\left(\text{smoothmax}\{T_{\text{miss},i} : t_{miss,i} \in d; k = 10\}, 0.1\right) \tag{25}$$

$$T_n\left[d\right] \sim \mathcal{N}\left(\text{smoothmax}\{T_{\text{miss},i} : t_{miss,i} \in d; k = 10\}, 0.1\right) \tag{26}$$

$$\tag{27}$$

Example imputations from this procedure are shown in Figure X. From May 25, 2015 to June 3, 2015, hourly temperatures are imputed at Waterloo Airport, using the hourly temperature measurements from nearby stations to inform the course of the temperatures, and using the daily minima and maxima "measurements" to constrain the imputed temperatures, and to infer the mean. Because we actually have hourly data for Waterloo, yet only fed our algorithm a reduction of this data to daily extremes, we can also plot the hidden temperatures, and see how faithfully the imputations reproduce them. We see that the imputations indeed track the true measurements very closely. The error bars satisfyingly narrow and widen in accordance to the amount of information available at each moments. On May 27th, we can see that the imputations capture the fact that the T_x record *could* have been set early in the measurement window, but more likely at its very end.

5 Model diagnostic

5.1 Variogram

We can visually inspect our model by plotting temporal and spatial semi-variograms. The semi-variogram of a stationary spatio-temporal function $Y(\mathbf{x}, t)$ is a function of the spatial lag $\mathbf{h}$ and the temporal lag r

$$\gamma\left(\mathbf{h}, r\right) = \frac{1}{2}\,\mathbb{E}\left[\left(Y\left(\mathbf{x}, t\right) - Y\left(\mathbf{x} + \mathbf{h}, t + r\right)\right)^2\right] = \text{var}\left(Y\left(\mathbf{x}, t\right)\right) - \text{cov}\left(\left(Y\left(\mathbf{x}, t\right)\right), Y\left(\mathbf{x} + \mathbf{h}, t + r\right)\right)\,. \tag{28}$$

For a Gaussian Process model, with a stationary kernel $k(\mathbf{h}, r) = k(\mathbf{x}, \mathbf{x} + \mathbf{h}, t, t + r)$ this can be expressed in terms of the observation noise σ_ϵ and $k(\cdot, \cdot)$, as

$$\gamma\left(\mathbf{h}, r\right) = \sigma_\epsilon^2 + k\left(0, 0\right) - k(\mathbf{h}, r)\,. \tag{29}$$

Furthermore, the semi-variogram can be estimated empirically, be averaging the square differences of any two observations that are separated by $\mathbf{h}$ in space, and r in time (or, in practice, within half a bin width of $\mathbf{h}$ and r). By comparing the empirical variogram to the variogram of our fitted $\mathcal{GP}$ model, we obtain a visual diagnosis of the model.

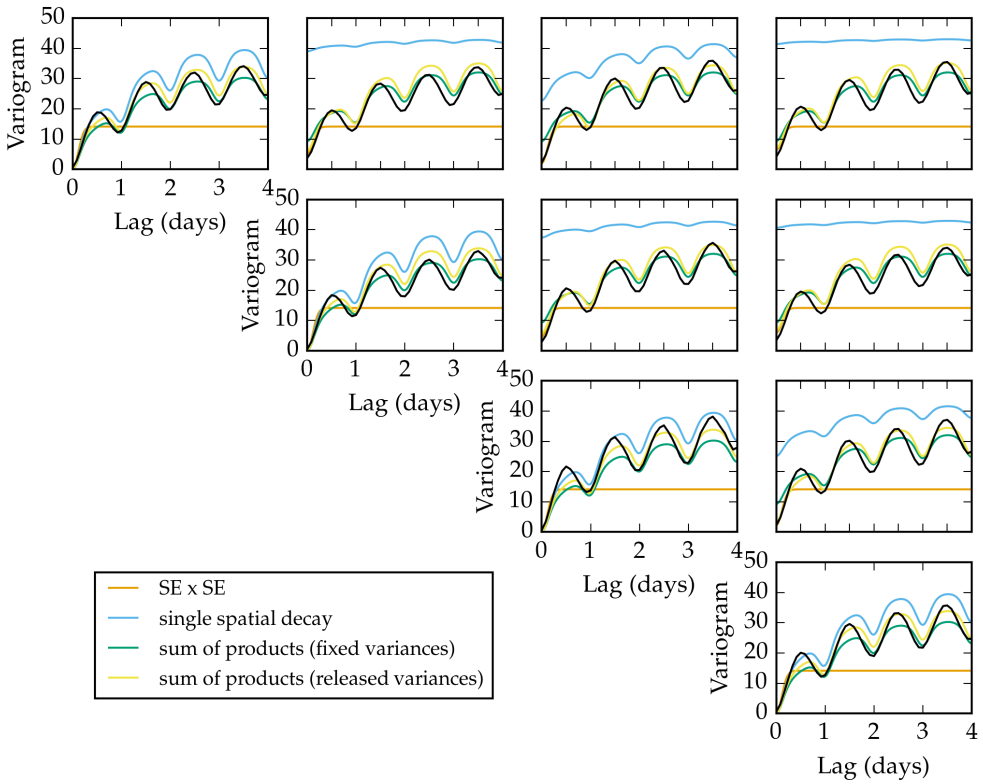


Figure 3: Semi-variogram

In our Iowa example, there are only four possible locations. For each location, we plot the empirical temporal variogram $\hat{\gamma}(0, r)$. For any pair of stations separated by $\mathbf{h}$ (fixed), we can also plot $\hat{\gamma}(\mathbf{h}, r)$. We then overlay the model's semi-variogram obtained through equation (29), resulting in Figure~(XX).

5.2 Error and expected error

In the previous section, we laid out our imputation strategy, and we visually checked that the imputations appear reasonable.

To quantify the predictive accuracy, we compare the posterior mean temperature to the truth, and obtain the empirical mean squared error as

$$\text{MSE}\left(\text{err} \mid T_{obs}, T_x, T_n\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}\left(T_{miss,i} \mid T_{obs}, T_x, T_n\right) - T_{miss,i}\right]^2 . \tag{30}$$

This equation is for the final predictions obtained using nearby hourly temperatures and local daily maxima and minima. A similar diagnostic can be computed for the intermediary predictions, excluding the local T_x and T_n. At that stage, we are not concerned with any overall bias in the predicted temperatures, so we instead obtain on the sample variance of the errors as

$$\text{var}\left(\text{err} \mid T_{obs}\right) = \text{var}_i\left\{\mathbb{E}\left(T_{miss,i} \mid T_{obs}\right) - T_{miss,i}\right\} . \tag{31}$$

Model	Likelihood	Var(err)	E(Var(err))	MSE(err)	E(MSE(err))
SE x SE	55,614	1.565	0.855	1.435	0.680
SExSE + diurnal	54,472	1.414	0.959	1.275	0.747

Model	Likelihood	Var(err)	E(Var(err))	MSE(err)	E(MSE(err))
Sum of products	47,082	1.244	2.031	~1.181	~0.959
SoP, fixed variance	48,589	3.813	7.619		
(time sum) x SE	52,467	1.439	17.996		

For our purposes, it isn't sufficient for the spatio-temporal model to yield good predictions; we also require a good estimate of its own accuracy. We estimate the expected MSE and predictive variance sampling K random draws T^k_{miss} from the posterior distribution, again conditioned on T_{obs} after fitting the spatio-temporal Gaussian process model, and on T_{obs}, T_x and T_n after incorporating the local data using Stan. The draws are from a multivariate normal distribution in the first case, and the MCMC samples obtained through Stan in the second case. We then evaluate the variance between the samples and the posterior mean as

$$\mathbb{E}\left(\mathrm{var}\left(\mathrm{err} \mid \cdot\right)\right) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{var}_i\left\{T^{(k)}_{miss,i} - \mathbb{E}\left(T_{miss,i} \mid \cdot\right)\right\} \tag{32}$$

When evaluating models, we want the errors to be small, and so the empirical error variance and MSE to be low. A well-calibrated model should also have the expected error variances $\mathbb{E}\left(\mathrm{var}\left(\mathrm{err} \mid \cdot\right)\right)$ close to their empirical values.

These diagnostics for our first spatio-temporal model, the product of squared exponentials, are found in the first row of Table X. The empirical error variance using only nearby measurements is already low, with typical errors of order $\sqrt{1.565} = 1.25\,^\circ$C. Incorporating the local measurements reduces it further to $\sqrt{1.435} = 1.20\,^\circ$C

6 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

7 Analysis

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

8 References

Baker, Donald G. (June 1975). "Effect of Observation Time on Mean Temperature Estimation". Journal of Applied Meteorology. 14 (4): 471–476.

-> follow up on citations to above