Experiments with Latent Precipitation Model

Tom Dietterich

V2 5 September 2022

Changes begin on slide 19

Fitting a Kriging model to the Oklahoma data

- Data: 366 days from 2008 OK Mesonet. 117 stations.
- Removed observations with QA flags indicating bad data. Most days there were only 116 stations with good data.
- Removed days if there was no rain anywhere in the state
- Normalized the precipitation numbers each day so that the maximum reported value is 1.0

Two Model-Fitting Problems

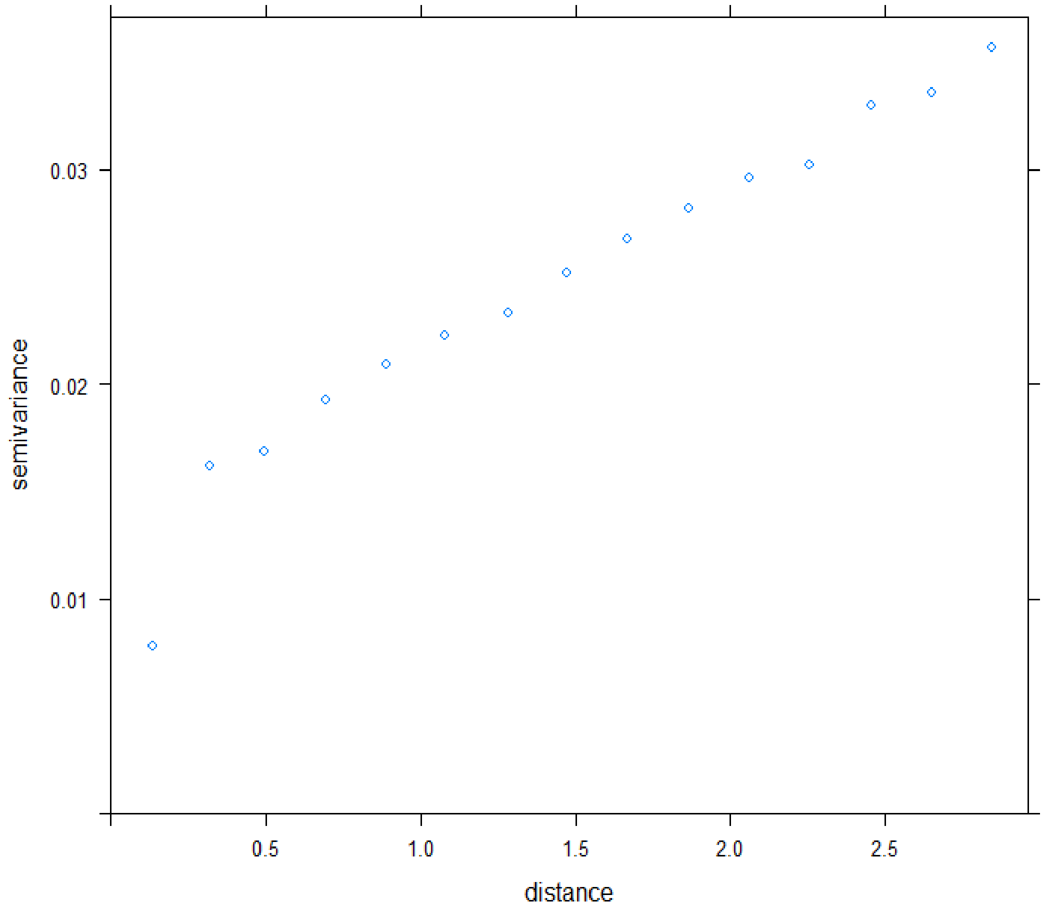
- Problem 1: Fit a spatial Gaussian process (“Kriging”) model. This involves fitting a semivariogram model γ, which defines the kernel for the GP
  - In Oklahoma, we have well-cleaned data, so I was able to do this.
  - In Africa, we could use manually-cleaned data from TAHMO
- Problem 2: For each day, we want to estimate the potential rain ψ at each station
  - For stations that reported $r_i > 0$, we define $\psi(x_i) = r_i$ and $d_i = 1$
  - For stations that reported $r_i = 0$, this could be due to a detection failure ($d_i = 0$) with $\psi(x_i) > 0$ or it could be a true zero with $\psi(x_i) = 0$
  - We are particularly interested in the $\psi(x_i) > 0$ and $d_i = 0$, because that is a “training example” for estimating θ_i, probability that $d_i = 1$
  - In Problem 2, we assume we know γ

Solving Problem 1:
Step 1: The Empirical Semivariogram

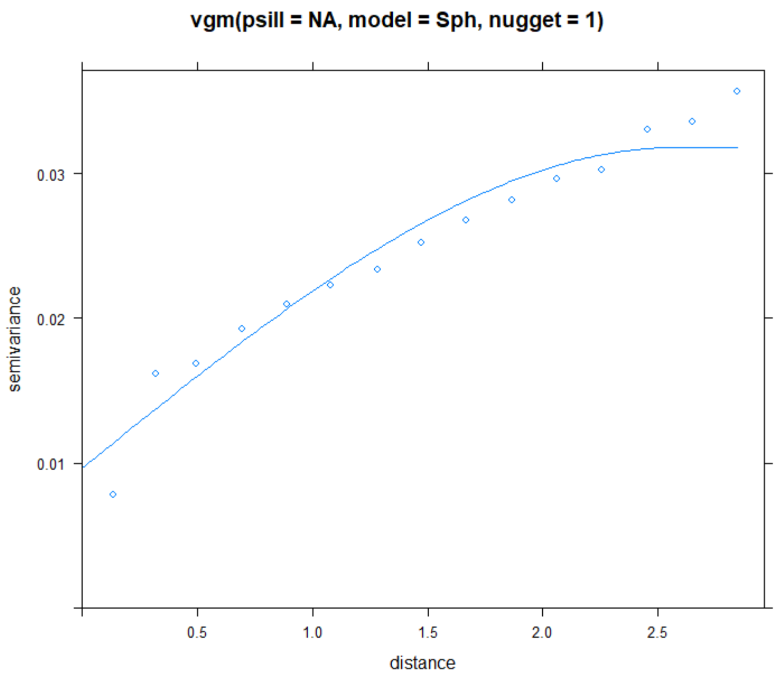
- R gstat package
- Pooling data across all (non-zero) days

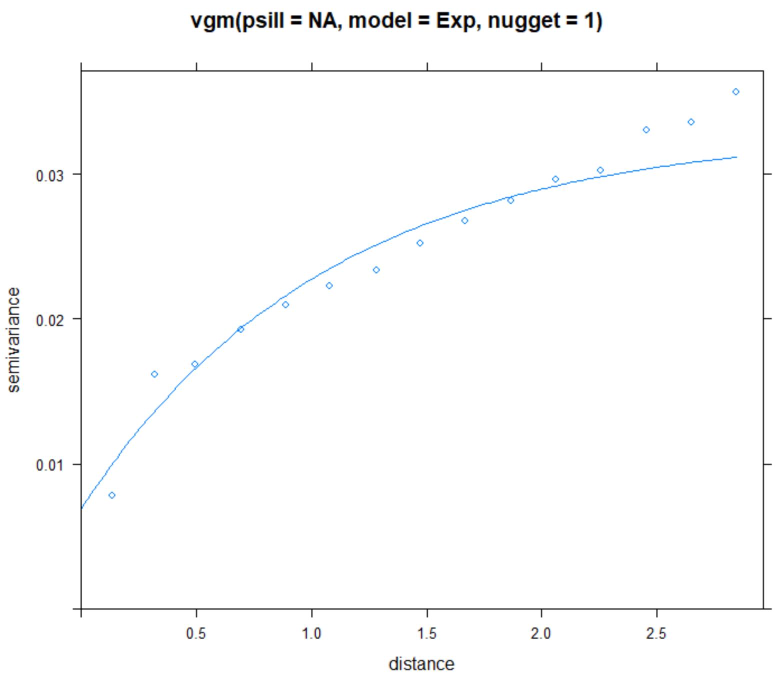
```
vg <- variogram(rain ~ day, data = df.year.scaled, dX = 0)
```

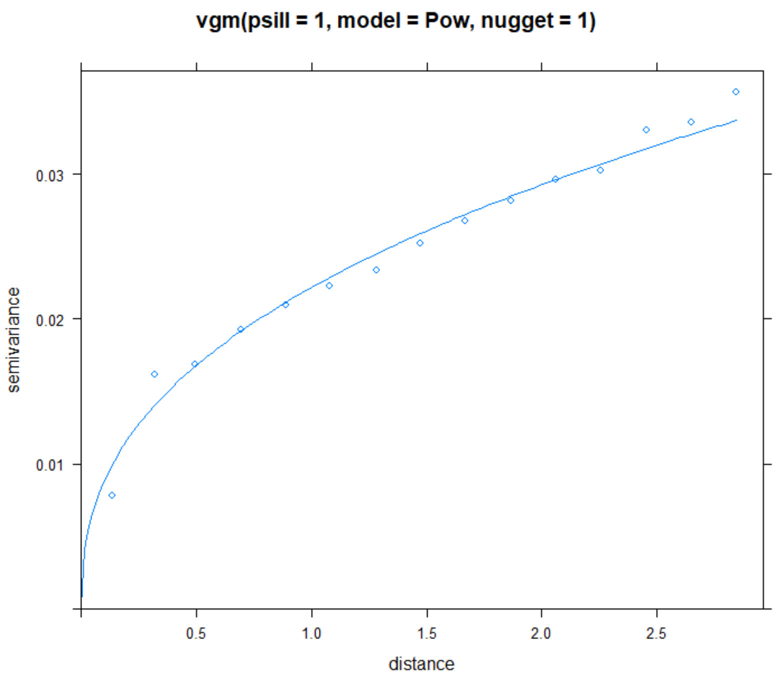
- The "dX=0" ensures that only station-days from the same day are compared to each other
- The variance is quite small and does not level off (no "sill"). This is a sign that there is probably a global trend. My guess is that this reflects the fact that from NW to SE the amount and frequency of rain increases
- Distance is in degrees of latitude or longitude. A better approach would be to choose a local map projection and then use km



Step 2: Variogram Model Selection



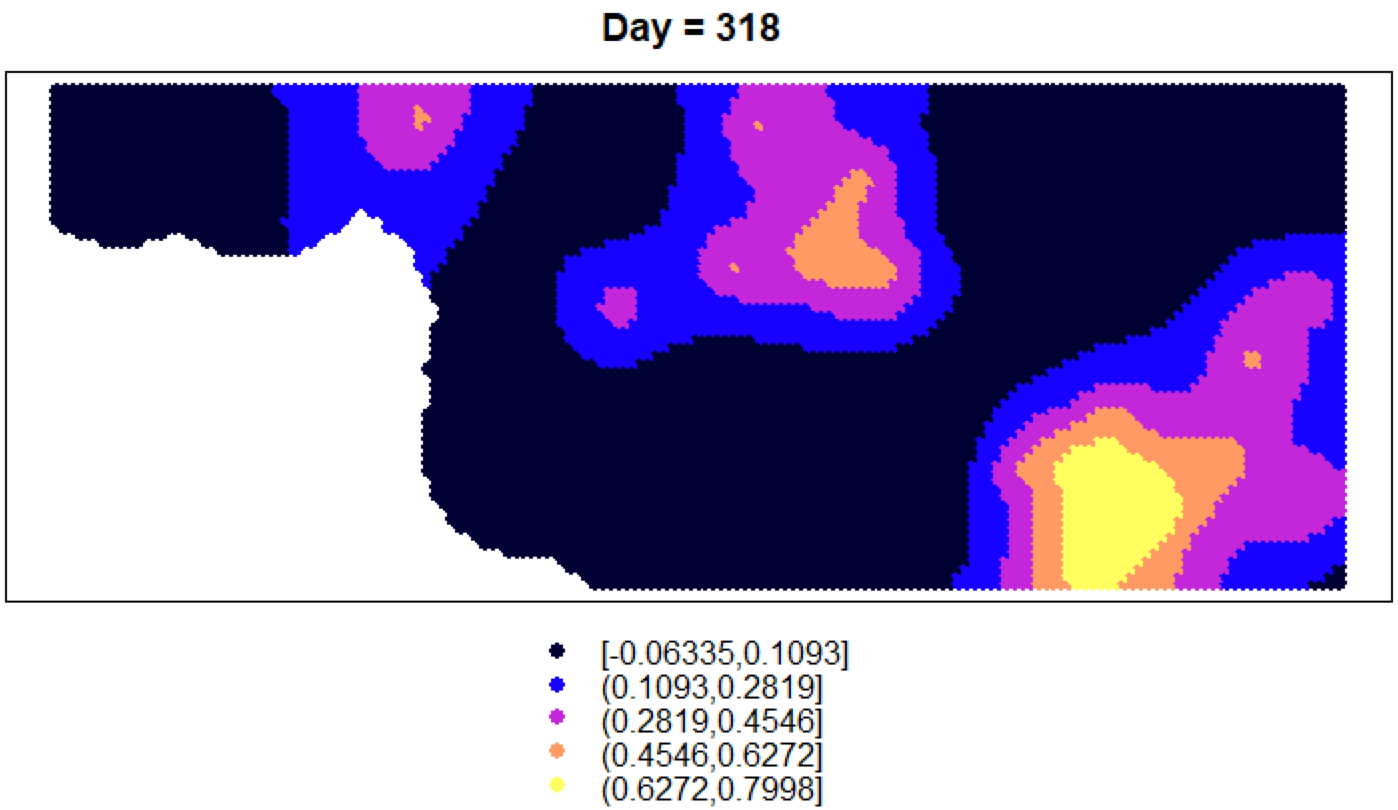




The psill parameter had no effect. The Pow model seemed to fit the best, although it has no nugget effect whereas both the Sph and Exp models include the nugget. But the rapid rise approximates a nugget very well, I think.

I did not include any measurement noise

Example Fitted Day



Problem 2: Estimating $\psi(x_i)$ at stations where $r_i = 0$ on a given day

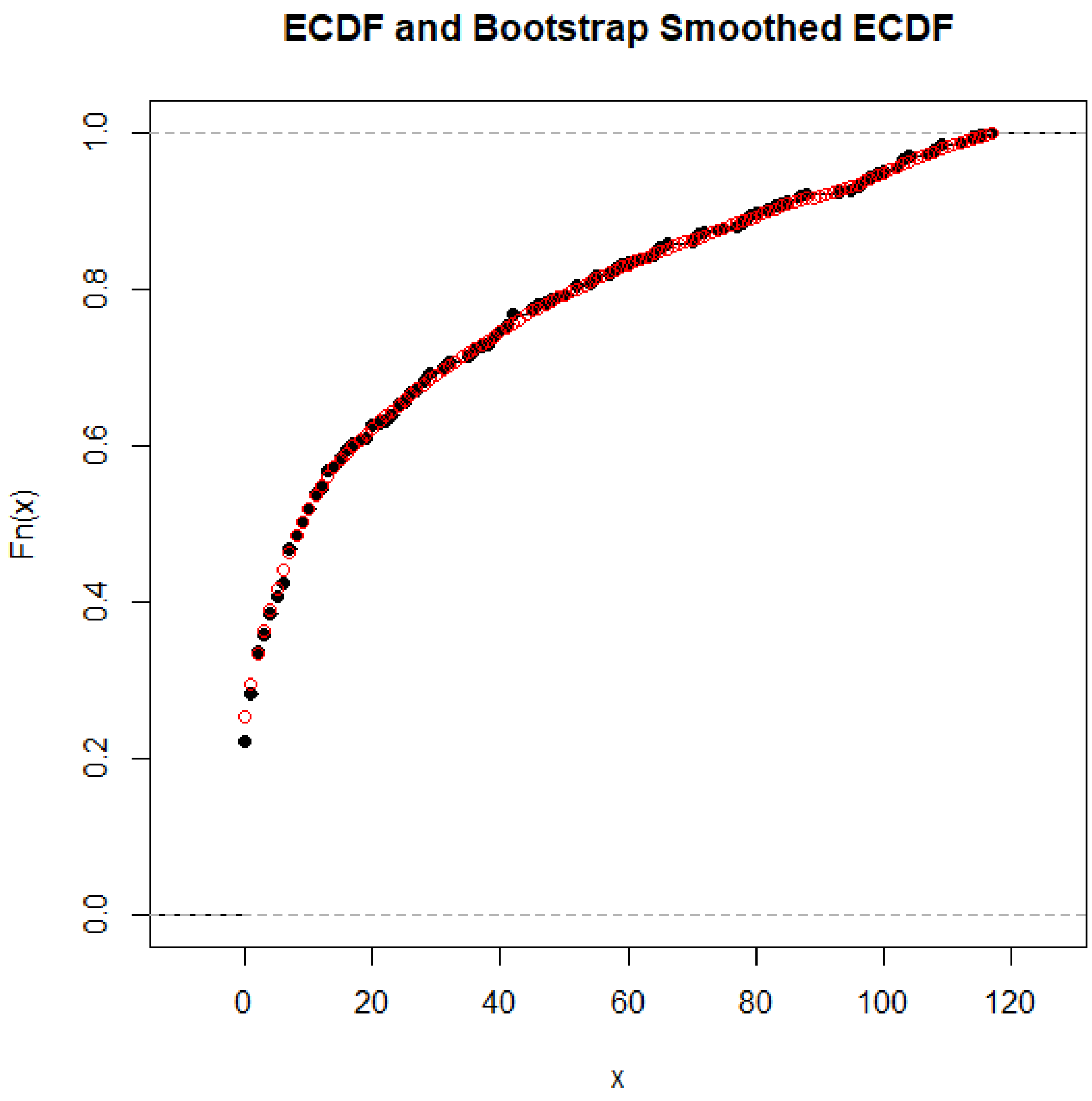
- Let $\tilde{R} = (\tilde{r}_1, \dots, \tilde{r}_n)$ be the reported rain values
- Let $R = (r_1, \dots, r_n)$ be the true ψ values
- Let $(d_1, \dots, d_n)$ be the binary detection variables
- Let $\theta_1, \dots, \theta_n$ be the Bernoulli parameters. On each day, $d_i \sim \mathrm{Bern}(d_i|\theta_i)$
- Let C be a candidate set of stations where we hypothesize that $\psi(x_i) > 0$ but $d_i = 0$. All other stations are hypothesized to either have $\psi(x_i) = 0$ or $d_i = 1$ (or both).
- We wish to estimate $P\left(C \middle| \tilde{R}\right)$

Probabilistic Model

- You might think that we could just simulate observations from the fitted Gaussian Process using γ. But I don't think this will work well, because the GP doesn't know how much rain to generate each day (because it doesn't model the mean of the GP).
- Proposed model:
  - Let $\#R$ be the number of stations with $r_i > 0$: $|\{i : r_i > 0\}|$
  - Let $\#\tilde{R}$ be the number of stations with $\tilde{r}_i > 0$
  - Estimate $P\left(\#R \middle| \#\tilde{R}\right)$
    - Fit $P(\#R)$ to the training data
    - Given a value for $\theta_i = \theta$, we know that $P(\#\tilde{R}|\#R)$ has a binomial distribution where we draw $\#R$ Bernoulli variables each with probability θ of being 1 and sum the values. Hence,
    - $P\left(\#R \middle| \#\tilde{R}\right) = \frac{1}{Z} P\left(\#\tilde{R} \middle| \#R\right) P(\#R)$ where $Z = \sum_{\#r} P\left(\#\tilde{R} \middle| \#R = \#r\right) P(\#R = \#r)$
  - Given C, we know that $\#R = \#\tilde{R} + |C|$ and each station in C must have $\psi > 0$. Hence, the probability is
  - $P\left(C \middle| \tilde{R}\right) = P\left(|C| + \#\tilde{R} \middle| \#\tilde{R}\right) P\left(C > 0 \middle| \tilde{R} \setminus C\right)$
    - This second probability can be computed from the GP

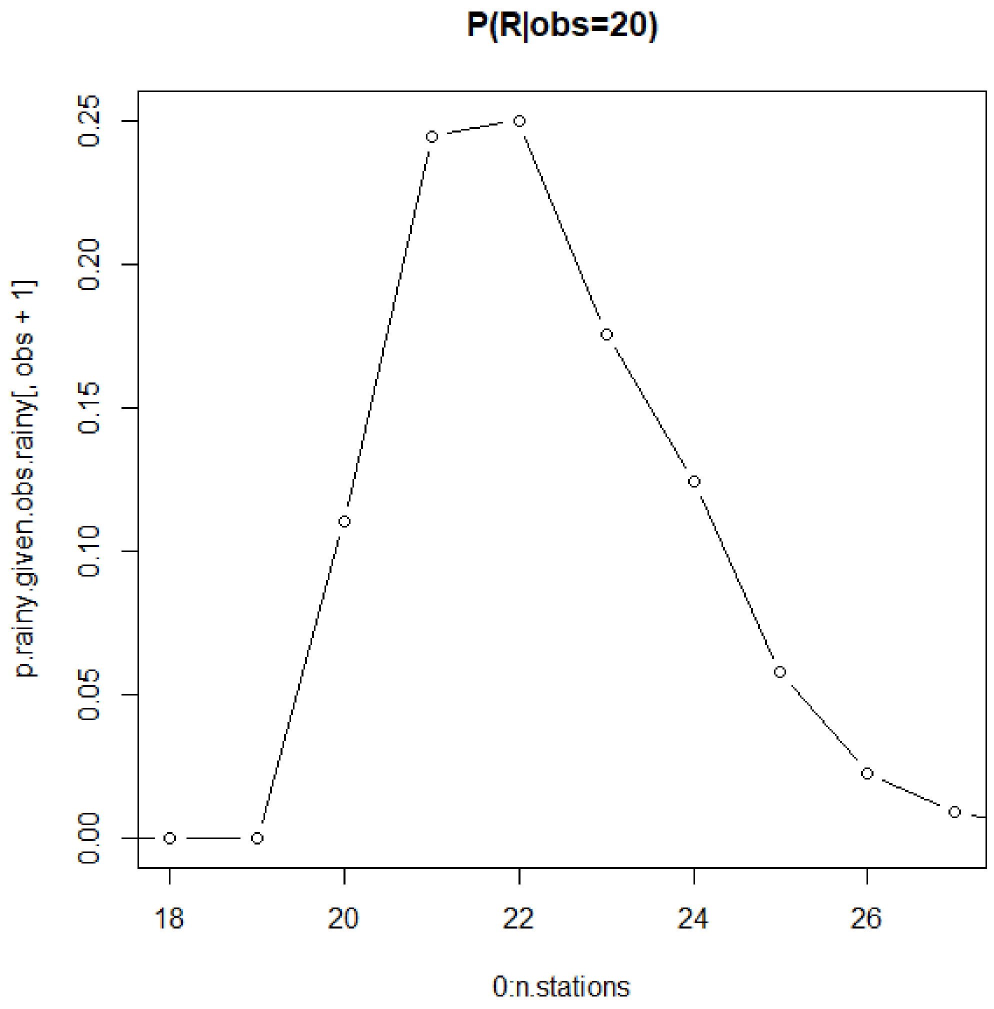
Fitting $P(\#R)$

- Using the 366 days, draw 100 bootstrap samples and compute $P(\#R)$ by pooling all of them. This smooths out the naïve estimate of $P(\#R)$ that we could get from just the 366 days
- The black points are the empirical CDF for $P(\#R)$ and the red points are the bootstrap version. We can see it has interpolated days where there where $\#R = x$ had no observations.



Computing $P\left(\#R \middle| \#\tilde{R}\right)$

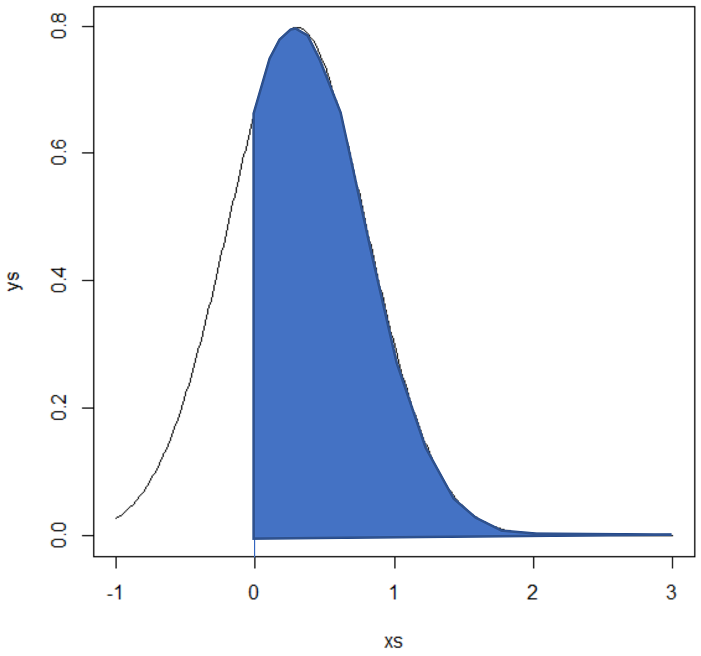
- I assumed a fixed $\theta = 0.9$ for all stations. A weakness of the model is that it requires a fixed θ.
- Here is a typical case. If $\#\tilde{R} = 20$ then the most likely value for $\#R$ is 21 or 22 but $\#R$ could be 20 or 23, 24, 25, 26. Larger values are highly unlikely



Computing $P\left(C > 0 \middle| \tilde{R} \setminus C\right)$

- $\tilde{R}$ is the observed rain
- $\tilde{R} \setminus C$ is the observed rain except at the stations in C
- Kriging gives us the predicted mean and variance of r_i for stations in C conditioned on the r_i values for $i \in \tilde{R} \setminus C$
    - We could request this as a multivariate Gaussian, but I treated the results as if the covariance matrix was diagonal
- $P\left(c_i > 0 \middle| \tilde{R} \setminus C\right)$ is the right tail measured from 0
- $P\left(C \middle| \tilde{R} \setminus C\right) = \prod_{i \in C} P\left(c_i \geq 0 \middle| \tilde{R} \setminus C\right)$
- Therefore:

$$P\left(C \middle| \tilde{R}\right) = P\left(|C| + \#\tilde{R} \middle| \#\tilde{R}\right) \prod_{i \in C} P\left(c_i \geq 0 \middle| \tilde{R} \setminus C\right)$$



Estimating ψ at the stations with $\tilde{r} = 0$

- Two approximation algorithms
  - Greedy construction of the single MLE C
  - Depth-first search of all “interesting” C candidate sets

Algorithm 1: Greedy MLE Method

- Let $Z = \{i | \tilde{r}_i = 0\}$
- $C \coloneqq \emptyset$
- $\ell \coloneqq P\left(C \middle| \tilde{R}\right)$
- repeat
  - $i^* \coloneqq \arg\max_i P\left(C \cup \{i\} \middle| \tilde{R} \setminus (C \cup \{i\})\right)$
  - Let $\ell^* \coloneqq P\left(C \cup \{i^*\} \middle| \tilde{R} \setminus (C \cup \{i^*\})\right)$
  - If $\ell^* < \ell$ return C
  - $C \coloneqq C \cup \{i^*\}$

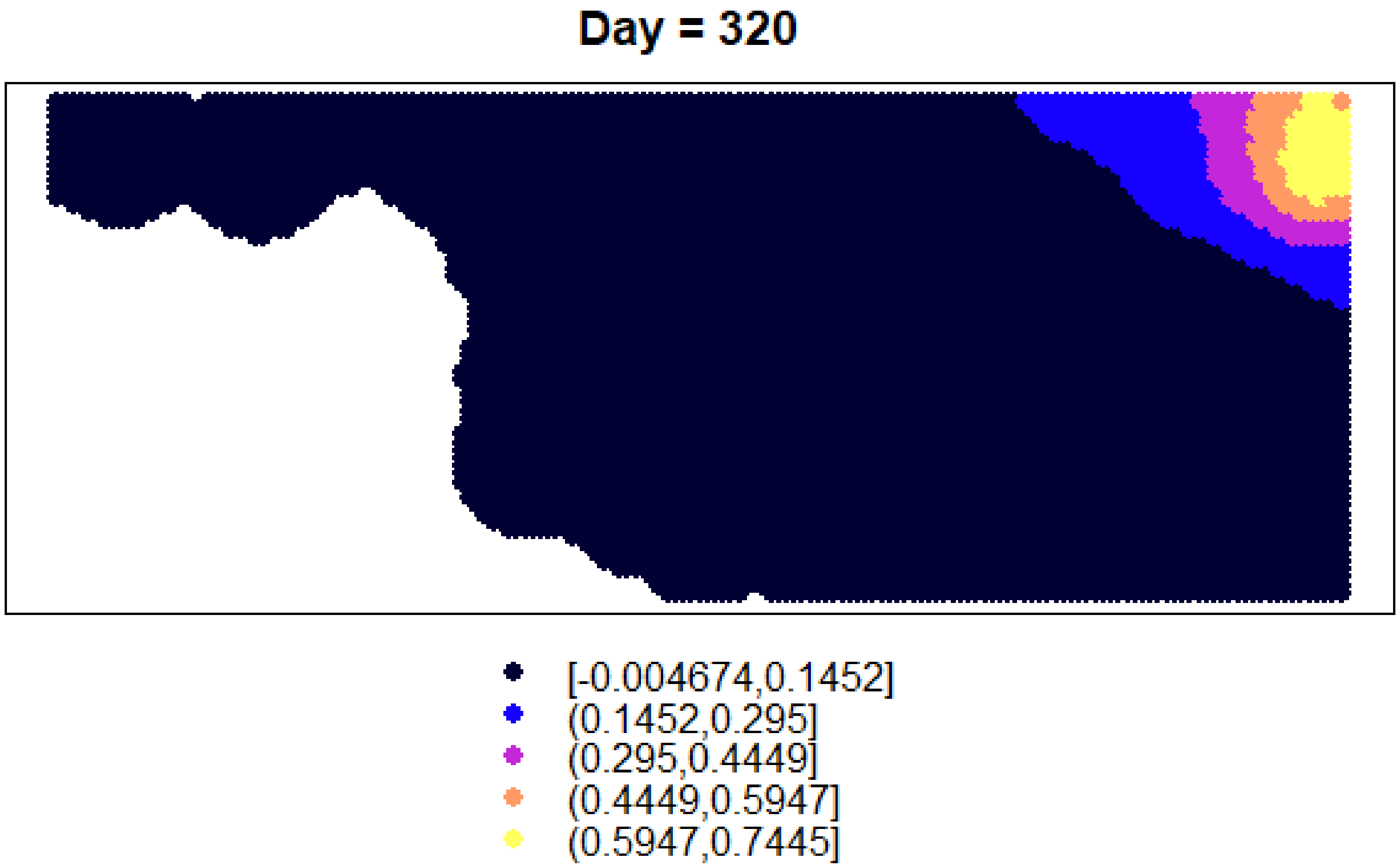
Greedy Method Results

- On days with a small amount of rain, the greedy method severely underestimates C
  - In 10 trials on day 318, it never found the correct stations
    - average true C contained 1.4 stations
    - computed C was always a single (incorrect) station 83
- On days with a large amount of rain, the greedy method often found the exact answer
  - In 10 trials on day 316, it found an average of 9.4 correct stations and missed an average of 2.0 stations. It was exactly correct in 5 trials

Algorithm 2: Depth-First Search of all Candidates

- The goal is to estimate for each station $i \in Z$, the probability that $r_i > 0$
- Plan:
  - Let $\mathcal{C}$ be the set of all possible candidate sets C
  - Compute $P\left(C \middle| \tilde{R} \setminus C\right)$ for each one
  - For each station i, $P\left(r_i > 0 \middle| \tilde{R}\right) \propto \sum_C \mathbb{I}[i \in C] P\left(C \middle| \tilde{R} \setminus C\right)$
- Problem: There are $2^{|Z|}$ possible candidate sets
- Solution: Depth-First Search with a likelihood cutoff
  - Let ℓ_{min} be the minimum likelihood of a candidate for it to be retained
  - $DFS(C)$:
    - Compute $\ell \coloneqq P\left(C \middle| \tilde{R} \setminus C\right)$
    - If $\ell > \ell_{min}$ add C to $\mathcal{C}$
    - If $P\left(|C| + \#\tilde{R} \middle| \#\tilde{R}\right) \times \max_{\#R} P\left(\#R \middle| \#\tilde{R}\right) > \ell_{min}$
      - Let $j^* \coloneqq \max_j j \in C$ be the highest-numbered station in C
      - For $j \in \{j^* + 1, \dots, |Z|\}$ $DFS(C \cup \{j\})$
  - Start by invoking $DFS(\emptyset)$

Performance of DFS method on day 320



Performance of DFS method on day 320

Correct $C = \{10, 19, 53, 66\}$

Top 10 Candidates

Threshold	# candidates	1	2	3	4	5	6	7	8	9	10
0.1	98	**53**	85	**66**	59	80	**19**	95	45	12	94
0.05	7090	**53**	85	**66**	80	59	**19**	45	95	12	94
0.025	129,494	**53**	85	**66**	59	**19**	95	45	12	94	69

Probability estimates

Threshold	# candidates	1	2	3	4	5	6	7	8	9	10
0.1	98	0.93	0.07	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01
0.05	7090	0.40	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
0.025	129,494	0.24	0.07	0.07	0.06	0.05	0.05	0.05	0.05	0.05	0.05

Greedy MLE = {**53**,85}

Station 10 is consistently missed by both methods

Summary of MLE estimation experiments

- Greedy MLE is not very good
- DFS with a high threshold is not too bad but its probability estimates are poor

Gaussian Noise Approximation

- Instead of the Bernoulli “noise” model, we can fit the GP with a Gaussian noise model.
  - The fitted model no longer exactly interpolates the data at the given stations
  - Therefore, at stations where $r = 0$, $\psi(x_i)$ may be > 0 because of nearby stations that reported $r > 0$

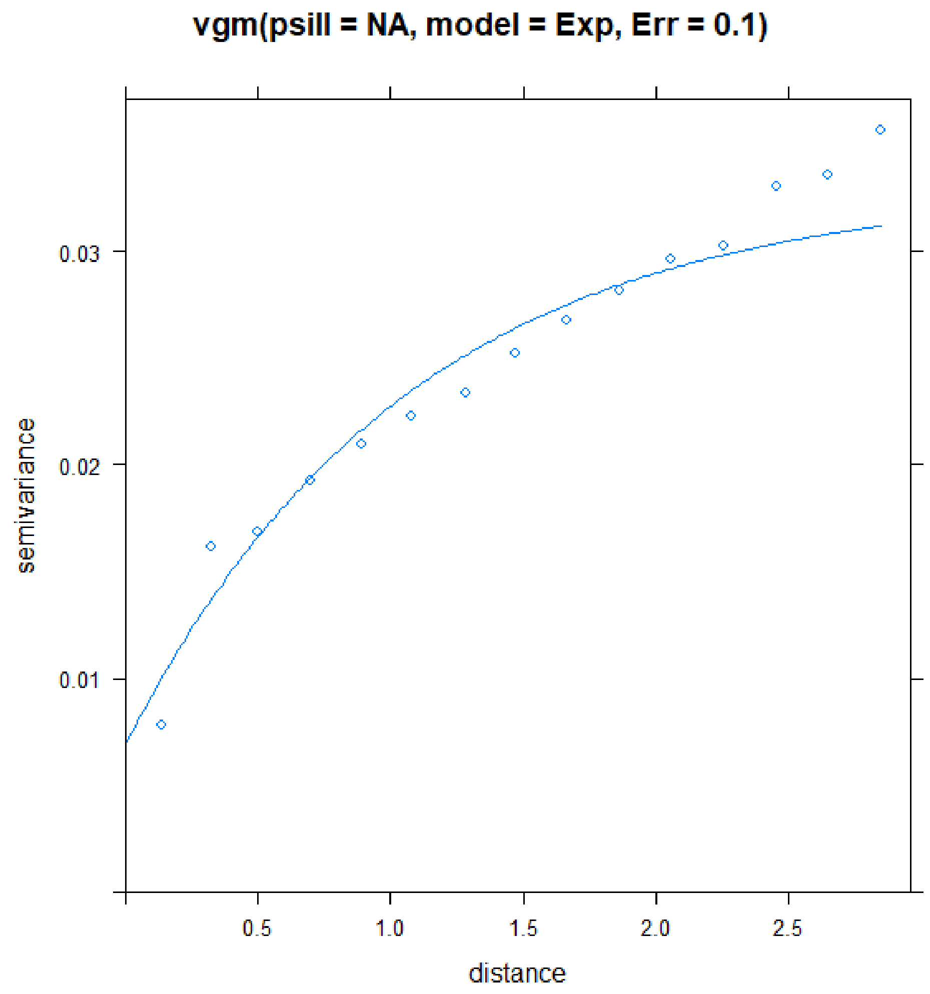
Variogram Models with Noise

- The gstat variogram model vgm(psill = 1, model = Pow, nugget = 1) is not compatible with a noise parameter. However, the Exp and Sph models do allow a noise parameter if you remove the nugget argument
- I chose the model
  - vgm(psill = 1, model = "Exp", Err = error.level)
  - The table at right shows the RMS error between r and $\hat{r}$ (predicted by Kriging)
  - The amount of error specified in the model had no effect as long as it was nonzero

	error.level	rms
1	0.0	3.423625e-15
2	0.1	1.807769e+00
3	0.2	1.807763e+00
4	0.3	1.807755e+00
5	0.4	1.807746e+00
6	0.5	1.807739e+00
7	0.6	1.807733e+00
8	0.7	1.807728e+00
9	0.8	1.807724e+00
10	0.9	1.806990e+00

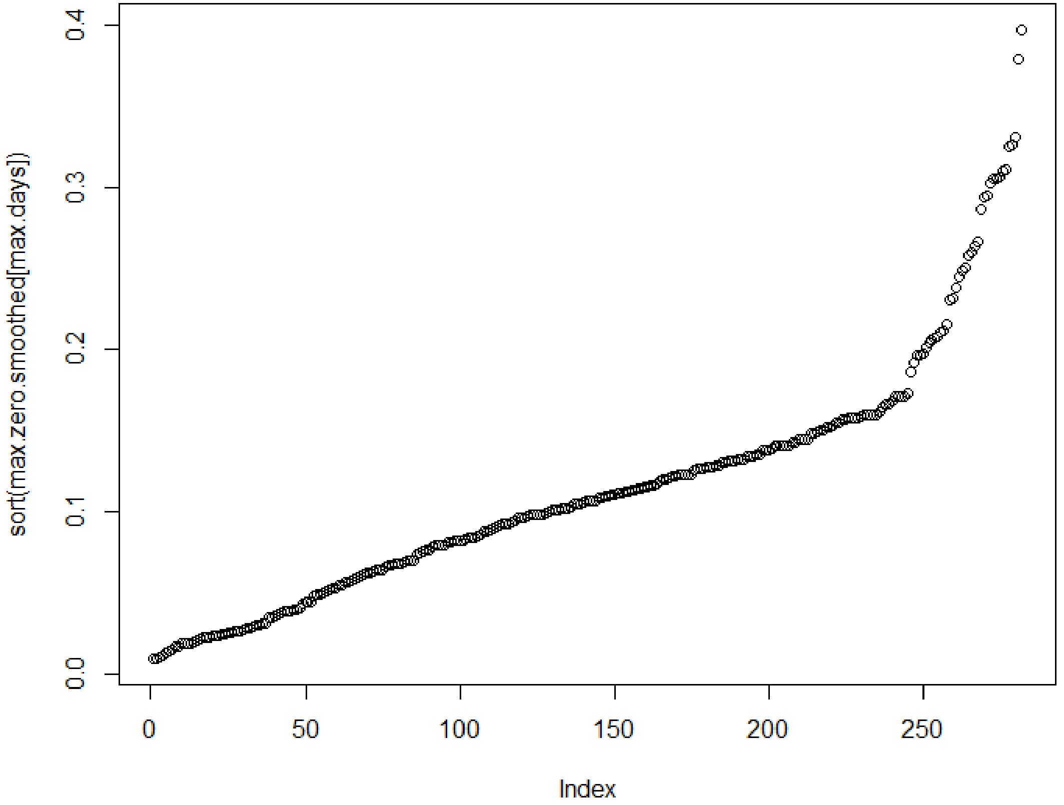
Chosen Variogram Model

- Adding the Err term makes no visible change in the variogram model
- But when applied to make predictions with kriging, it no longer interpolates the data points



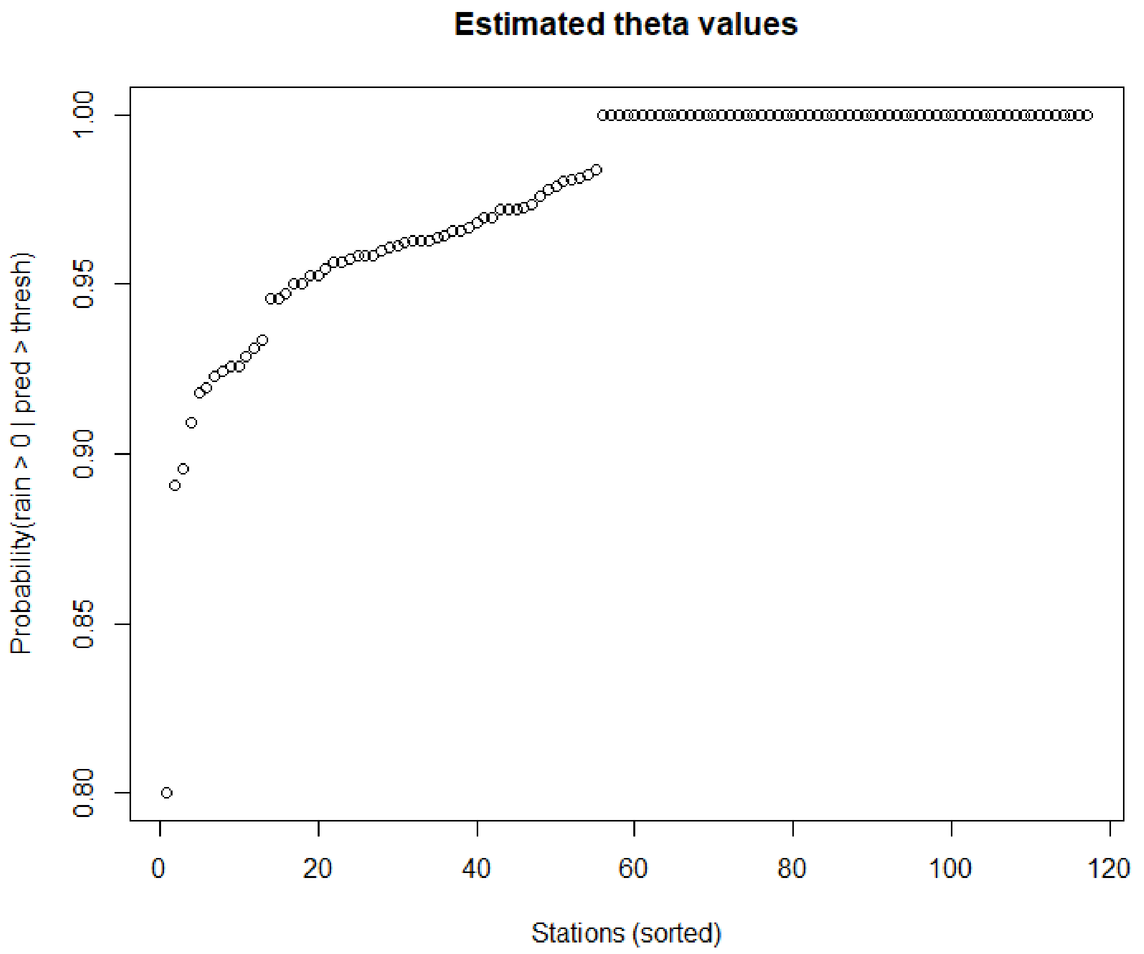
Choosing the zero threshold ϵ

- The predicted rain $\hat{r}$ is never exactly zero, so we need to define a threshold
- For each day with total rain > 0, we compute the maximum value of $\hat{r}$ at any station reporting $r = 0$
- If we set the threshold at 0.4, we would get no "false nonzero" cases. However, because of local heterogeneity of precipitation, we expect to have a certain number of such false nonzero cases (where $r = 0$ but $\hat{r} > \epsilon$, our threshold)
- We chose $\epsilon = 0.18$, as this is right at the "elbow" where there seems to be some regime change



Baseline θ estimates

- With this value for ϵ and no simulated blocked sensors, we can estimate the θ values
- One station seems to have quite a low value



CUSUM Detection Statistic

- Cortese (2015) PhD thesis “Change Point Detection and Estimation in Sequences of Dependent Random Variables” gives a very nice review of CUSUM methods for Boolean random variables
- Given a sequence of Bernoulli random variables $x_1, \dots, x_n$
  - H_0: All variables share the same Bernoulli parameter p (no change point)
  - H_a: $x_1, \dots, x_t$ have parameter θ_1 and $x_{t+1}, \dots, x_n$ have parameter θ_2
- Uncorrected CUSUM statistic compares the running total to the expected running total (based on H_0) at each time t:

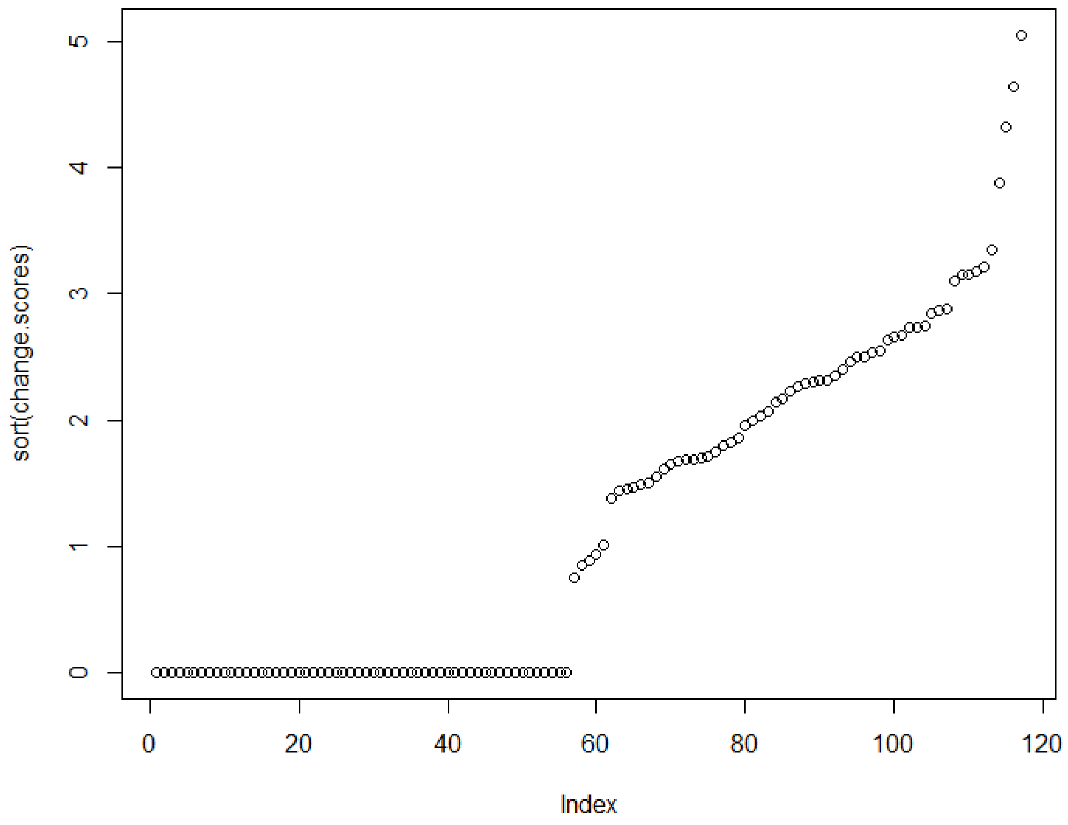
$$S_t = \sum_{j=1}^{t} x_j - \frac{t}{n} \sum_{j=1}^{n} x_j = \sum_{j=1}^{n} a_j x_j, \quad \text{where} \quad a_j = \begin{cases} 1 - \frac{t}{n} & \text{if } 1 \le j \le t, \\ -\frac{t}{n} & \text{if } t+1 \le j \le n. \end{cases}$$

Standardizing the CUSUM statistic

- The standard deviation of the CUSUM statistic under H_0 is
  - $\hat{\sigma}_t = \sqrt{p(1-p)\frac{t}{n}\left(1-\frac{t}{n}\right)}$
  - It is larger in the middle of the sequence and smaller at the ends
- The standardized CUSUM statistic is
  - $T_t = \frac{CUMSUM_t}{\hat{\sigma}_t} = \frac{S_t}{\hat{\sigma}_t\sqrt{n}}$
- Standardizing allows us to compare T_t values across all $t = 1, \ldots, n$
- The asymptotic sampling distribution under the null is known, but we will choose a significance cutoff empirically

Distribution of T_t under the null

- For our 117 stations, for Oklahoma data from 2008 (285 rainy days), we obtain the following distribution of T_t
- This suggests choosing a threshold Δ around 3.3-3.5 which will give us 4 false alarms even without inserting any faults

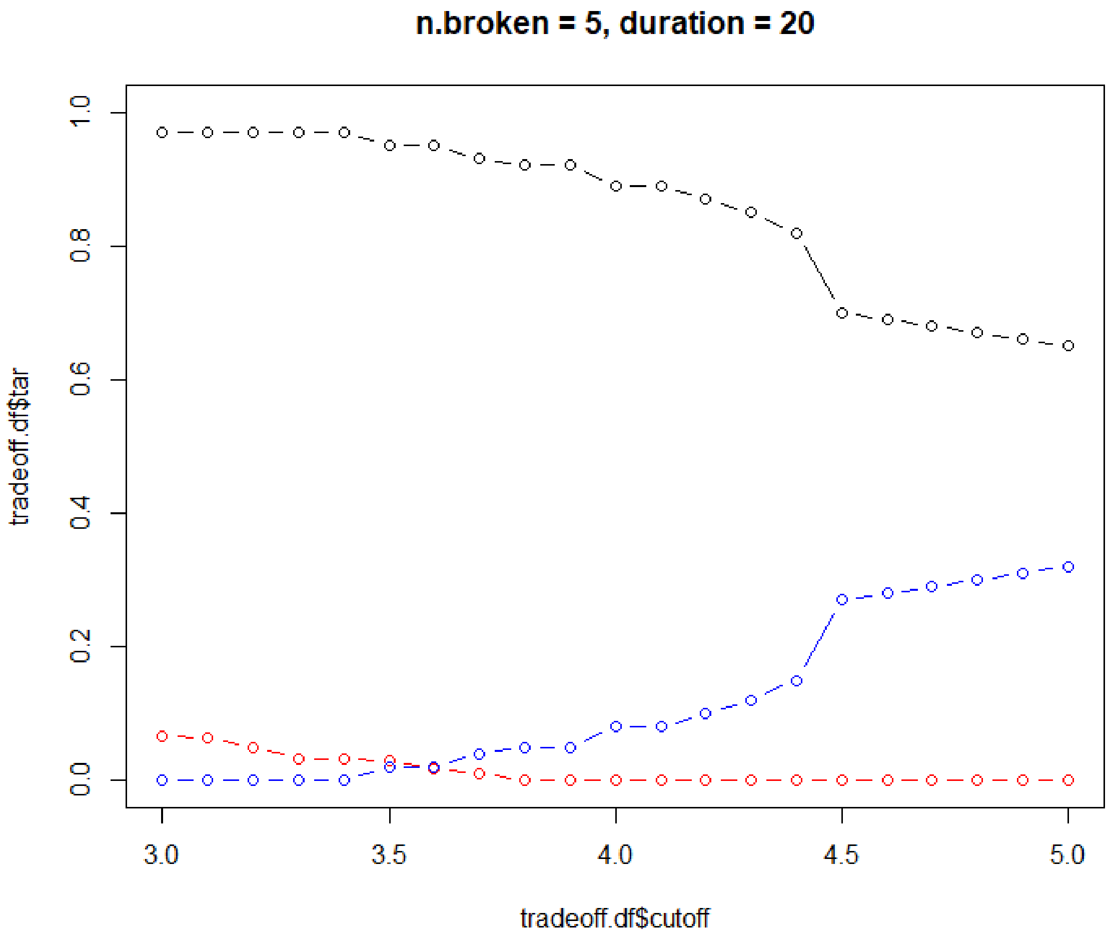


Methodology for Simulating Blocked Sensors

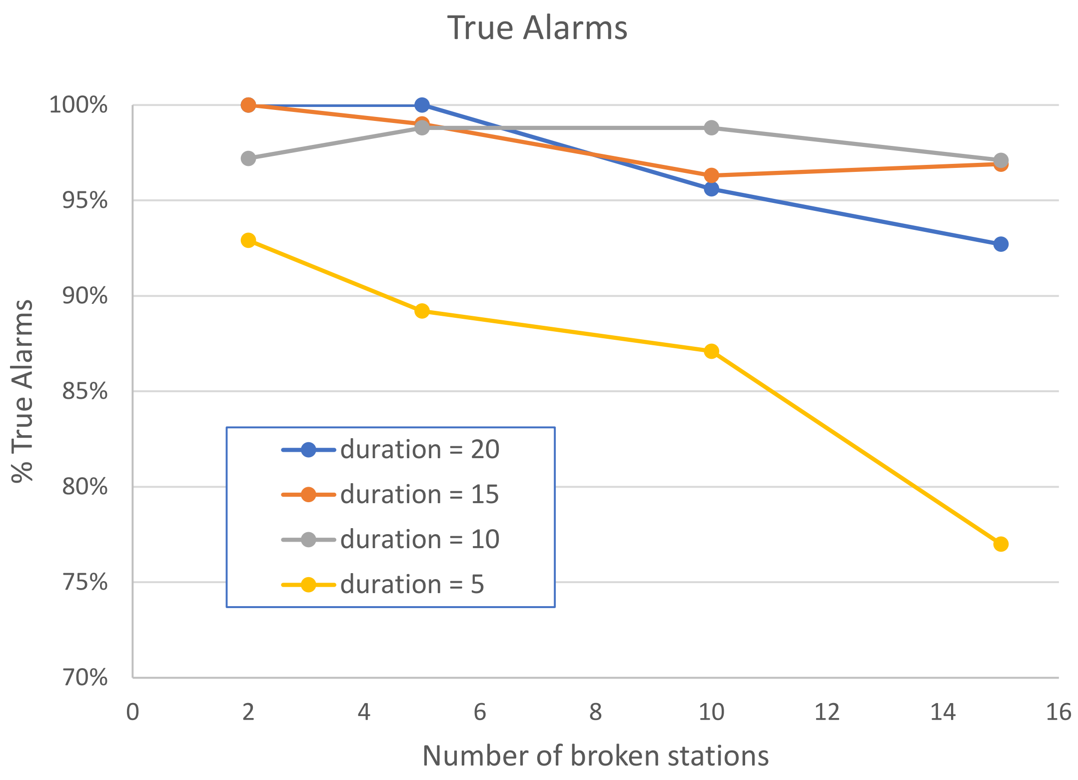
- Choose B stations to simulate a blockage
- For each station $b \in \{1, \dots, B\}$, choose a time t such that there are m days $j \in [t, 366]$ with $r_{b,j} > 0$
- Replace those values with zero
  - An additional condition on t is that there be at least 10 days $j \in [1, t-1]$ with $r_{b,j} > 0$ so that the shift from $r_{b,j} > 0$ to $r_{b,j} = 0$ can be detected
- For each rainy day of the year
  - Fit the kriging model and compute $\hat{r}_{i,j} = \psi(i, j)$ for each station i and day j
  - If $\hat{r}_{i,j} \geq 0.18$ and $r_{i,j} = 0$, then $d_{i,j} = 0$ else $d_{i,j} = 1$
  - Note that because of the inserted zeros, the number of values where $d_{b,j} = 0$ is usually much less than m. For example, we might use $m = 20$ but only observe between 2 and 7 days with $d_{b,j} = 0$
- Compute $T_{i,t}$ for $t \in [1{,}366]$ and let T_i^* be the largest value (and t_i^* be the corresponding time)
- If $T_i^* > \Delta$, then declare station i to be blocked starting at time t_i^*
  - We also require $\sum_{t=t^*}^{n} d_{i,t} \geq 2$. This eliminates many false alarms without introducing any missed alarms
  - It requires a minimum post-change-point sample size of 2

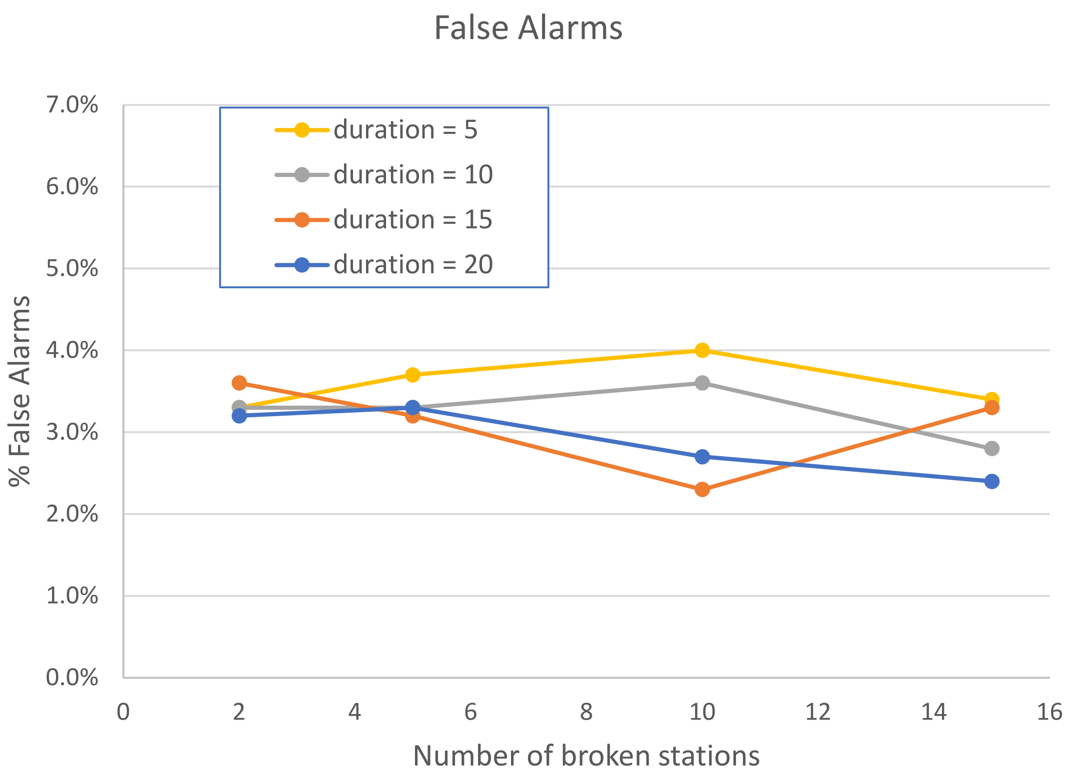
Choosing Δ

- Run 20 replicates with $B = 5$ stations blocked with $m = 20$ nonzero readings and aggregate all of the T_i^* scores
- Vary Δ and compute the true alarm rate (black), false alarm rate (red), and missed alarm rate (blue)
- Selected $\Delta = 3.4$
  - We do not want to miss broken sensors
  - This is the largest value that has zero missed alarms



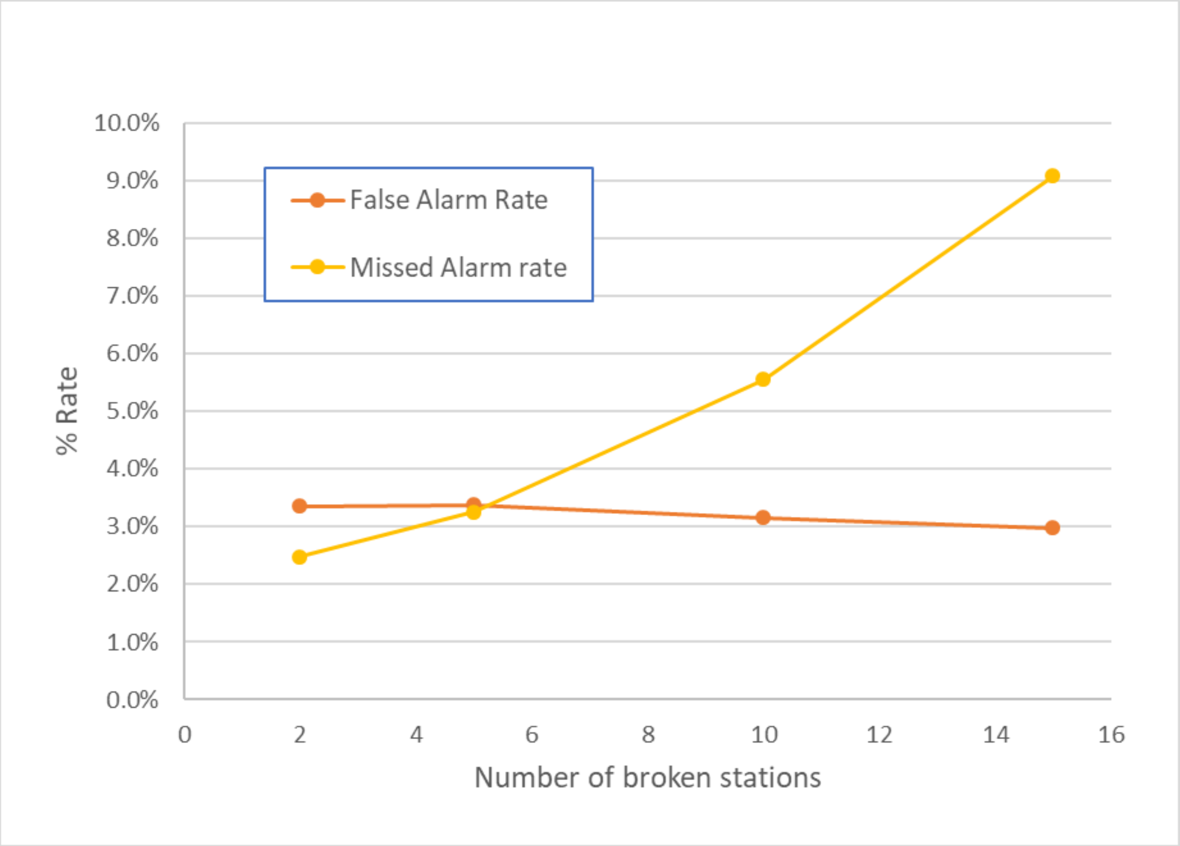
True Alarms and False Alarms





Aggregated Experiment Summary

- False Alarm Rate is basically constant at 3%
- Missed Alarm Rate increases as we break more stations
  - More broken stations cause ψ to be zero in more other stations
  - This prevents us from inferring d and detecting the blockage
  - 15/117 = 12.8% of stations blocked
  - We are detecting >90% of blocked stations



Summary

- This method is simple and very promising
  - I think we should test it on TAHMO data
  - I hope we can deploy it
  - Deployment will require lots of additional research and engineering
- This method is a hack
  - It would be nice to have a more elegant solution
    - Cirra has a variational idea he is studying
  - Why doesn't the Err parameter in the variogram model affect the smoothness of the fitted model?