Spatial Analysis of Crowdsourced Mobile Data

Arnab Chakraborty, Soumendra Lahiri, Alyson Wilson

Department of Statistics
North Carolina State University

May 18, 2018

1. Introduction
2. Standard Approach
3. Veracity Scores
4. VS-based Methodology
5. Monte Carlo Study
6. Case Study
7. Takeaway

1. Introduction
2. Standard Approach
3. Veracity Scores
4. VS-based Methodology
5. Monte Carlo Study
6. Case Study
7. Takeaway

Introduction

- **Crowdsourced Mobile Data:** Data on geographic elements such as ambient temperature etc., captured by sensors installed in mobile devices and gathered by mobile applications e.g. *AccuWeather*, *WeatherSignal* etc.

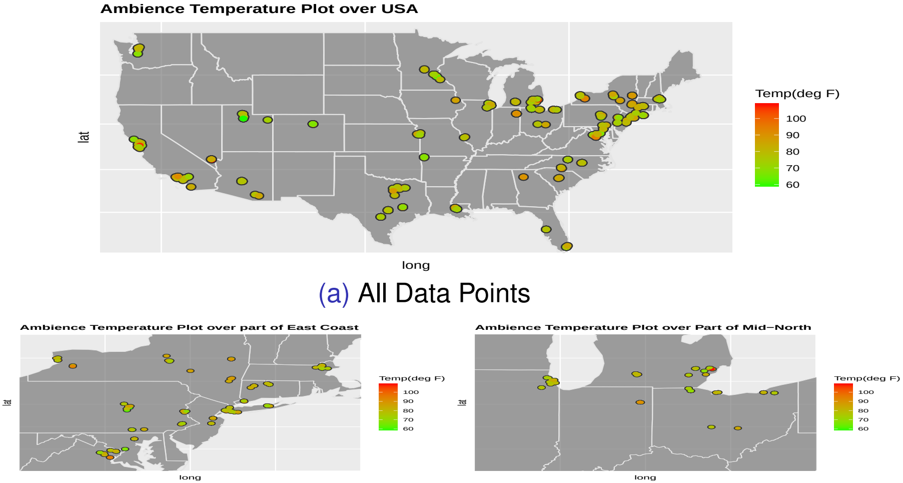
Introduction

- A **potential data source** for ‘hyper-local’ analysis of weather elements, disaster detection, specially in regions with *less ground stations* and *high population densities.*
- Due to the omnipresence of mobile devices global leaders in weather information technology, e.g. AccuWeather, OpenSignal etc. are turning each app-user to a weather station.
- But the ‘amature’ quality of the sensors, the non-laboratory environment affects the reliability of the crowdsourced mobile data making the analysis challenging.

Data Description

- Dataset for this project is gathered by ***WeatherSignal***, a mobile application by *OpenSignal*, to gather crowdsourced weather information from mobile devices.
- For the course of this study we are interested in Daily Average Ambient Temperature process for a particular day (04/30/13) over the land of the USA.
- The observations coming from mobile sensors have varying quality and we believe an unknown portion of the data is contaminated due to interaction with unknown processes.

Spatial Plots of The Data

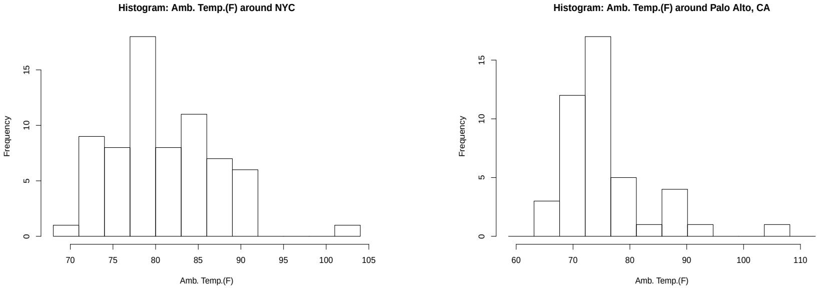

Ambience Temperature Plot over USA
lat
long
Temp(deg F)
100
90
80
70
60
Ambience Temperature Plot over part of East Coast
lat
long
Temp(deg F)
100
90
80
70
60
Ambience Temperature Plot over Part of Mid−North
lat
long
Temp(deg F)
100
90
80
70
60

(a) All Data Points

(b) Upper East

(c) Mid North

Histogram of Observations



(a)

(b)

Figure: Empirical distribution of average temperatures in two regions: (a) New York City (b) Palo Alto, CA, with area 0.2° × 0.1°.

Histogram of Observations

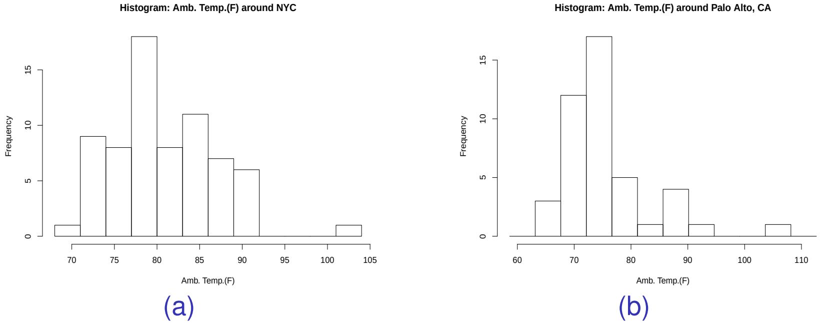


Figure: Empirical distribution of average temperatures in two regions: (a) New York City (b) Palo Alto, CA, with area $0.2^\circ \times 0.1^\circ$.

Assessment of data reliability needed!

Objective

- Assessment of data quality/reliability: continuous scoring.
- Incorporation of the score to develop a robust technology for analysis of spatial data.
- Evaluation of the new robust approach as compared to the standard methodology.
- Spatial interpolation of the process over a fine resolution grid using both ground station measured data and crowdsourced information.

1. Introduction
2. **Standard Approach**
3. Veracity Scores
4. VS-based Methodology
5. Monte Carlo Study
6. Case Study
7. Takeaway

Model

- Let $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2\}$ be a real-valued spatial process observed at finite number of irregularly spaced locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$ in $\mathcal{D}$; $\mathbf{Z} = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$.
- $Z(\mathbf{s})$ is assumed to have a decomposition of the form,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D},$$

  where $E(Y(\mathbf{s})) = \mu(\mathbf{s})$; $\epsilon(\mathbf{s})$ is a spatially correlated mean-zero random “error" process.
- Under spatial regression setup, $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\beta$; $\mathbf{x}(\cdot) = (x_1(\cdot), ..., x_p(\cdot))'$ is known vector of covariates.
- Assumption: $\epsilon(\mathbf{s})$ is intrinsically stationary with variogram $2\gamma(\mathbf{h}; \boldsymbol{\theta}) = \text{var}\{\epsilon(\mathbf{s}) - \epsilon(\mathbf{s} + \mathbf{h})\}$, $\boldsymbol{\theta}$ is the covariance parameter.

Standard Approach

- Estimate the mean parameters, $\hat{\beta}_{OLS} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \{Z(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)'\beta\}^2 = (X'X)^{-1} X'\mathbf{Z}$.
- Covariance parameters are estimated by least squares-based variogram model fitting from the observed residuals, $\hat{\epsilon} = \mathbf{Z} - \hat{\mathbf{Z}}$, as

$$\hat{\theta}_{WLS} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{j=1}^{k} w_j \left\{\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j; \boldsymbol{\theta})\right\}^2 .$$

- Ordinary kriging is used to interpolate the residual process over the space and the spatial prediction is obtained as

$$\hat{Z}(\mathbf{s}_0) = \mathbf{x}\left(\mathbf{s}_0\right)' \hat{\beta}_{OLS} + \hat{\epsilon}(\mathbf{s}_0).$$

1. Introduction
2. Standard Approach
3. Veracity Scores
4. VS-based Methodology
5. Monte Carlo Study
6. Case Study
7. Takeaway

Veracity Scores: Motivation

Example

$$Y(\mathbf{s}_i) = \mu + \epsilon(\mathbf{s}_i), \quad Z(\mathbf{s}_i) = \epsilon_{M_i} Y(\mathbf{s}_i) + \epsilon_{A_i},$$

where, $\{\epsilon_{M_i}\}_{i=1}^n \underset{\text{indep.}}{\sim} \left(1, \sigma^2_{M_i}\right)$, $\{\epsilon_{A_i}\}_{i=1}^n \underset{\text{indep.}}{\sim} \left(0, \sigma^2_{A_i}\right)$ & $\{\epsilon_{M_i}\}_{i=1}^n \in \mathbb{R}^+$.

Consider a weighted average, $\hat{\mu}_{\text{VS}} = \frac{\sum_{i=1}^n v(\mathbf{s}_i) Z(\mathbf{s}_i)}{\sum_{i=1}^n v(\mathbf{s}_i)}$. Then,

$$\mathrm{Var}\left(\hat{\mu}_{\text{VS}}\right) = \frac{\sigma^2_\epsilon \sum_{i=1}^n v(\mathbf{s}_i)^2}{\left(\sum_{i=1}^n v(\mathbf{s}_i)\right)^2} + \frac{\left(\mu^2 + \sigma^2_\epsilon\right) \sum_{i=1}^n v(\mathbf{s}_i)^2 \sigma^2_{M_i}}{\left(\sum_{i=1}^n v(\mathbf{s}_i)\right)^2} + \frac{\sum_{i=1}^n v(\mathbf{s}_i)^2 \sigma^2_{A_i}}{\left(\sum_{i=1}^n v(\mathbf{s}_i)\right)^2} + \frac{\sigma^2_\epsilon \sum\limits_{i_1 \neq i_2} v(\mathbf{s}_{i_1}) v(\mathbf{s}_{i_2}) \rho_\epsilon(\mathbf{s}_{i_1} - \mathbf{s}_{i_2})}{\left(\sum_{i=1}^n v(\mathbf{s}_i)\right)^2}.$$

Veracity Scores: Motivation

Illustration: Take $\sigma^2_{M_i} = \sigma^2_{A_i} = Ci^{\alpha}$ and $v(\mathbf{s}_i) = i^{-\beta}$, for some $\alpha \geq 0$ and $\beta \geq 0$.

Table: Variances of $\hat{\mu}$ ($\beta = 0$) and $\hat{\mu}_{\mathrm{VS}}$ ($\beta > 0$). The true parameters are taken to be : the population mean $\mu = 5$, residual variance $\sigma^2_{\epsilon} = 3$ and the spatial correlation parameter $\rho = 0.5$.

α	n	$\beta = 0$	$\beta = 0.5$	$\beta = 1$
	100	14.734	8.513	5.954
1	500	14.547	7.770	4.483
	1000	14.523	7.609	4.050
α	n	$\beta = 0$	$\beta = 0.5$	$\beta = 1$
	100	981.304	423.909	108.135
2	500	4847.861	1938.833	314.458
	1000	9681.180	3800.256	517.735

Veracity Scores: Formulation

Consider $\mathcal{B}_\delta(\mathbf{s}_i) = (\mathbf{s}_i - \delta, \mathbf{s}_i + \delta]$ containing $\mathbf{s}_i$. Let $\mathbf{Z}_i = \left(Z(\mathbf{s}_{i_1}), ..., Z(\mathbf{s}_{i_{n(i)}})\right)'$ be the data vector with locations $\in \mathcal{B}_\delta(\mathbf{s}_i)$.

Veracity Score (VS) of the observation at $\mathbf{s}_i$:

$$V(\mathbf{s}_i) = \phi\left(\frac{|Z(\mathbf{s}_i) - \xi(\mathbf{Z}_i)|}{\mathcal{R}(\mathbf{Z}_i)}\right),$$

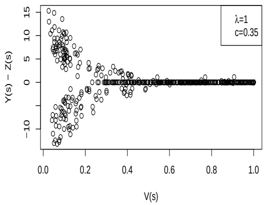
where,

- ◇ $\phi : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ is a function such that $\phi(x) \downarrow 0$ as $x \to \infty$ and $\phi(x) \leq \phi(0) < \infty$.
- ◇ $\xi(\mathbf{Z}_i)$, $\mathcal{R}(\mathbf{Z}_i)$ are some measures of central tendency and dispersion of the observations $\left\{Z(\mathbf{s}_{i_1}), ..., Z(\mathbf{s}_{i_{n(i)}})\right\}$.

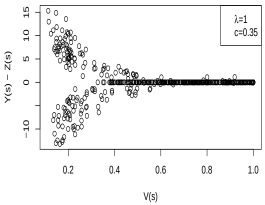
Veracity Scores: Illustration

- *Lower* values of VS indicate poor quality of the observation.
- Two variants of VS have been used in this study:
  1. **Mean-VS:** $\xi(\mathbf{Z}_i) = \bar{Z}_{i\cdot}$, and $\mathcal{R}(\mathbf{Z}_i) = \text{s.d.}(\mathbf{Z}_i)$.
  2. **Median-VS:** $\xi(\mathbf{Z}_i) = Q_2(\mathbf{Z}_i)$, and $\mathcal{R}(\mathbf{Z}_i) = IQR(\mathbf{Z}_i)$.
- Performance of VS on synthetic data:

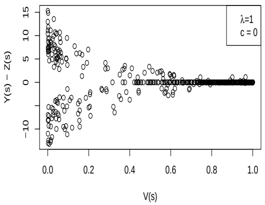
Polynomial Veracity Function



Logarithmic Veracity Function



Exponential Veracity Function



1 Introduction

2 Standard Approach

3 Veracity Scores

4 VS-based Methodology

5 Monte Carlo Study

6 Case Study

7 Takeaway

VS-based Model Fitting & Kriging

- $\hat{\beta}_{OLS} \longrightarrow \hat{\beta}_{VS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}\left(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\beta\right)$, where $\mathcal{L}(\cdot, \cdot)$ is some loss function.

VS-based Model Fitting & Kriging

- $\hat{\beta}_{OLS} \longrightarrow \hat{\beta}_{VS} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}\left(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\beta\right)$, where $\mathcal{L}(\cdot,\cdot)$ is some loss function.

- $\hat{\epsilon}(\mathbf{s}_i) \longrightarrow \tilde{\epsilon}(\mathbf{s}_i) = V(\mathbf{s}_i)^q \hat{\epsilon}(\mathbf{s}_i) + (1 - V(\mathbf{s}_i)^q) Q_2(\hat{\boldsymbol{\epsilon}}_i)$, where $\hat{\boldsymbol{\epsilon}}_i = \left(\hat{\epsilon}(\mathbf{s}_{i_1}), \ldots, \hat{\epsilon}(\mathbf{s}_{i_{n(i)}})\right)$.

VS-based Model Fitting & Kriging

- $\hat{\beta}_{OLS} \longrightarrow \hat{\beta}_{VS} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}\left(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\beta\right)$, where $\mathcal{L}\left(\cdot,\cdot\right)$ is some loss function.

- $\hat{\epsilon}(\mathbf{s}_i) \longrightarrow \tilde{\epsilon}(\mathbf{s}_i) = V(\mathbf{s}_i)^q \hat{\boldsymbol{\epsilon}}(\mathbf{s}_i) + (1 - V(\mathbf{s}_i)^q)Q_2(\hat{\boldsymbol{\epsilon}}_i)$, where $\hat{\boldsymbol{\epsilon}}_i = \left(\hat{\epsilon}(\mathbf{s}_{i_1}), \ldots, \hat{\epsilon}(\mathbf{s}_{i_{n(i)}})\right)$.

- $\hat{Z}(\mathbf{s}_0) \longrightarrow \hat{Z}_{VS}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\hat{\beta}_{VS} + \tilde{\epsilon}(\mathbf{s}_0)$.

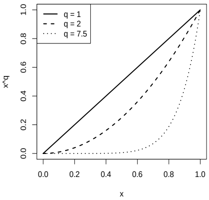
VS-based Model Fitting & Kriging

- $\hat{\beta}_{OLS} \longrightarrow \hat{\beta}_{VS} = \underset{\beta}{\text{argmin}} \; \sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}\left(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\beta\right)$, where $\mathcal{L}(\cdot, \cdot)$ is some loss function.

- $\hat{\epsilon}(\mathbf{s}_i) \longrightarrow \tilde{\epsilon}(\mathbf{s}_i) = V(\mathbf{s}_i)^q \hat{\epsilon}(\mathbf{s}_i) + (1 - V(\mathbf{s}_i)^q) Q_2(\hat{\boldsymbol{\epsilon}}_i)$, where $\hat{\boldsymbol{\epsilon}}_i = \left(\hat{\epsilon}(\mathbf{s}_{i_1}), \ldots, \hat{\epsilon}(\mathbf{s}_{i_{n(i)}})\right)$.

- $\hat{Z}(\mathbf{s}_0) \longrightarrow \hat{Z}_{VS}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\hat{\beta}_{VS} + \tilde{\epsilon}(\mathbf{s}_0)$.

- q is the parameter for regulating the degree of smoothing needed: chosen optimally to minimize cross-validated MAPE or RMSPE.



1 Introduction

2 Standard Approach

3 Veracity Scores

4 VS-based Methodology

5 Monte Carlo Study

6 Case Study

7 Takeaway

Simulation

- $Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\beta + \epsilon(\mathbf{s})$, $Z(\mathbf{s}_i) = \epsilon_{M_i} Y(\mathbf{s}_i) + \epsilon_{A_i}$.
- $\left(\epsilon_{M_i}, \epsilon_{A_i}\right)' \mid \left(\sigma^2_{M_i}, \sigma^2_{A_i}\right)' \underset{\text{indep.}}{\sim} \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{M_i} & 0 \\ 0 & \sigma^2_{A_i} \end{pmatrix}\right)$, and $\sigma^2_{M_i} \underset{\text{i.i.d.}}{\sim} \sigma^2_{0M} \times \text{Ber}(p_M)$ & $\sigma^2_{A_i} \underset{\text{i.i.d.}}{\sim} \sigma^2_{0A} \times \text{Ber}(p_A)$. Proportion of noisy observations: p_e.

Simulation

- $Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \epsilon(\mathbf{s})$, $Z(\mathbf{s}_i) = \epsilon_{M_i} Y(\mathbf{s}_i) + \epsilon_{A_i}$.
- $\left(\epsilon_{M_i}, \epsilon_{A_i}\right)' \mid \left(\sigma^2_{M_i}, \sigma^2_{A_i}\right)' \underset{\text{indep.}}{\sim} \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{M_i} & 0 \\ 0 & \sigma^2_{A_i} \end{pmatrix} \right)$, and $\sigma^2_{M_i} \underset{\text{i.i.d.}}{\sim} \sigma^2_{0M} \times \text{Ber}(p_M)$ & $\sigma^2_{A_i} \underset{\text{i.i.d.}}{\sim} \sigma^2_{0A} \times \text{Ber}(p_A)$. Proportion of noisy observations: p_e.

β_0	p_e	**Median VS**	**Mean VS**	**Std. App.**
	19%	2.532	2.175	1
55	36%	3.433	1.998	1
	51%	4.298	1.740	1
β_x	p_e	**Median VS**	**Mean VS**	**Std. App.**
	19%	2.520	2.289	1
1.5	36%	3.811	2.615	1
	51%	5.091	2.397	1

Table: Relative efficiencies

1 Introduction

2 Standard Approach

3 Veracity Scores

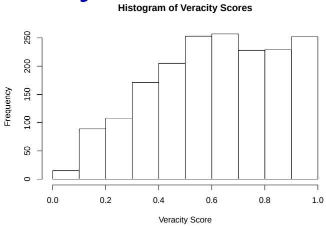
4 VS-based Methodology

5 Monte Carlo Study

6 Case Study

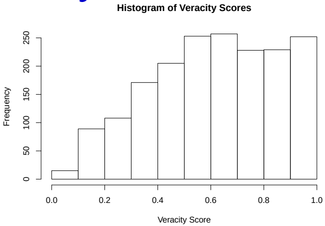
7 Takeaway

Case Study Details

- 

The histogram of the veracity scores for our data.

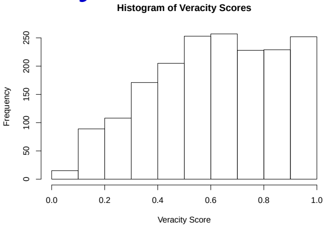
Case Study Details

- 

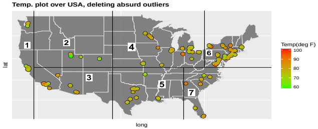
The histogram of the veracity scores for our data.

- Among $n = 1848$ observations there are about 350 observations with VS less than 0.4 indicating the noisy nature of the data.

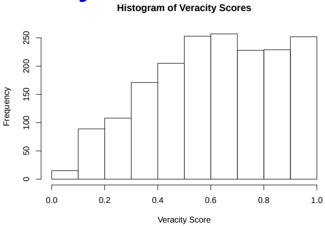
Case Study Details



- The histogram of the veracity scores for our data.
- Among $n = 1848$ observations there are about 350 observations with VS less than 0.4 indicating the noisy nature of the data.
- Instead of 'global' model we have fitted regional models for the 7 blocks as shown in the picture.



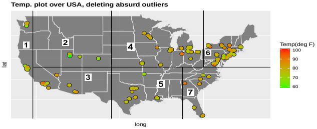
Case Study Details

- 

  The histogram of the veracity scores for our data.

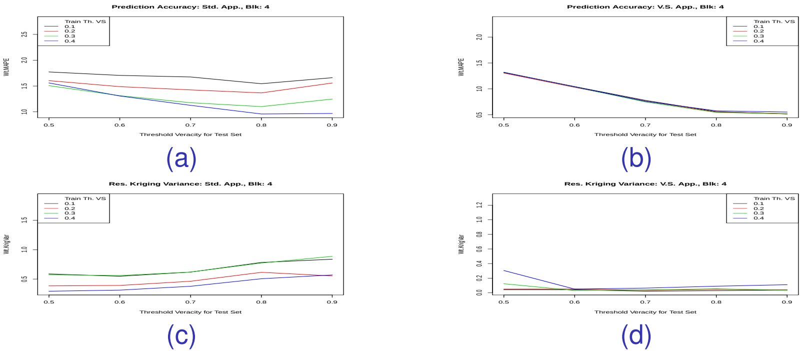
- Among $n = 1848$ observations there are about 350 observations with VS less than 0.4 indicating the noisy nature of the data.

- Instead of 'global' model we have fitted regional models for the 7 blocks as shown in the picture.



- We focused our analysis on the observations in block 4 and 6 as these regions has reasonable number of observations.

Comparison Analysis: Block 4



(a)

(b)

(c)

(d)

Figure: L1O-MAPE's and ARKV's for different combinations of training and test set threshold VS for both standard approach (left) and VS-based method for block 4.

Kriging & Imputation

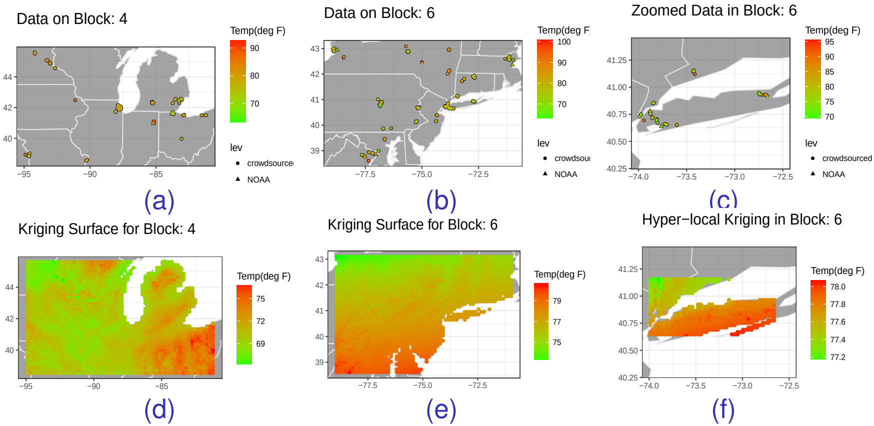


Figure: Kriging surfaces & imputation between ground stations using LM (d, e) & B-Spline (f) mean models and estimated covariance models.

1 Introduction

2 Standard Approach

3 Veracity Scores

4 VS-based Methodology

5 Monte Carlo Study

6 Case Study

7 Takeaway

Summary

- Introduced assessment of data quality through Veracity Score in spatial setting.
- Modified standard spatial analysis incorporating veracity score in mean and covariance structure estimation and kriging equation.
- Ground station temperatures along with the crowdsourced data are used for spatial imputation of daily average ambient temperature process.
- In future we will try to apply this method to real-time data by considering a neighborhood in both spatial and temporal dimension to define VS.
- For more details see: `https://ncsu-las.org/las-technical-reports/`.

References

- AccuWeather (2015), ‘Accuweather launches accucast, providing exclusive crowdsourced weather feature worldwide’, **URL:** `https://www.accuweather.com/en/ress/50601069`.
- Cressie, N. (1993), Statistics for spatial data, John Wiley & Sons, Inc.
- Cressie, N. & Douglas, H. M. (1980), ‘Robust estimation of variogram I.’, *Journal of the International Association for Mathematical Geology* **12**(2), 115-125.
- Sosko, S. & Dalyot, S. (2017), ‘Crowdsourcing user-generated mobile sensor weather data for densifying static geosensor networks’, *ISPRS International Journal of Geo-Information* **6**(3), 61.

Thank You