The Annals of Applied Statistics
2020, Vol. 14, No. 1, 116–142
https://doi.org/10.1214/19-AOAS1290

A STATISTICAL ANALYSIS OF NOISY CROWDSOURCED WEATHER DATA

BY ARNAB CHAKRABORTY*, SOUMENDRA NATH LAHIRI AND ALYSON WILSON

*Department of Statistics, North Carolina State University, *arnab2897@gmail.com*

Spatial prediction of weather elements like temperature, precipitation, and barometric pressure are generally based on satellite imagery or data collected at ground stations. None of these data provide information at a more granular or "hyperlocal" resolution. On the other hand, crowdsourced weather data, which are captured by sensors installed on mobile devices and gathered by weather-related mobile apps like `WeatherSignal` and `AccuWeather`, can serve as potential data sources for analyzing environmental processes at a hyperlocal resolution. However, due to the low quality of the sensors and the nonlaboratory environment, the quality of the observations in crowdsourced data is compromised. This paper describes methods to improve hyperlocal spatial prediction using this varying-quality, noisy crowdsourced information. We introduce a reliability metric, namely Veracity Score (VS), to assess the quality of the crowdsourced observations using a coarser, but high-quality, reference data. A VS-based methodology to analyze noisy spatial data is proposed and evaluated through extensive simulations. The merits of the proposed approach are illustrated through case studies analyzing crowdsourced daily average ambient temperature readings for one day in the contiguous United States.

1. Introduction.

In recent years there has been a proliferation of weather-related applications for mobile devices such as cellphones, iPods and laptops. These applications not only provide service to the user but also collect and share spatial data on location, ambient temperature, barometric pressure, humidity, etc., captured by the small-scale sensors installed in the devices. Analyzing and understanding these crowdsourced datasets is becoming an area of increasing interest.

One use of the mobile sensor-generated data is to analyze and understand atmospheric processes at very fine spatial resolution. Most of the methodologies in literature for spatial prediction of weather elements are based on global images coming from satellites or measurements taken at meteorological stations on the ground (e.g., see Thornton, Running and White (1997); Florio et al. (2004), etc.). But none of these sources are dense enough so that the variability of the process can be analyzed in "hyperlocal" regions, for example, rectangular regions inside the population centers with each sides varying approximately in between 25 to 30 miles ($0.3° - 0.6°$ in latitude and longitude). For instance, the ground stations are generally situated away from localities, for example, at airports or national parks, etc. Hence, weather-related analysis solely based on ground-station data does not often provide correct assessment of the variation of the underlying process in the localities. However, in disaster detection, traffic management and many defense-related activities, prediction of the process in a very localized region (hyperlocal) is often more important than the global imputation of the process over a bigger region. Crowdsourced data captured by mobile sensors can serve as a potential source in these scenarios, especially in regions where the ground weather stations are sparse but the population density, and hence the density of the mobile devices like cellphones, iPads, etc., is relatively high. Recently, a handful of organizations are becoming interested in providing cost-effective hyperlocal predictions of weather using sensor-generated geographical information through weather-related mobile apps. For example, the global leader

Received March 2019; revised June 2019.

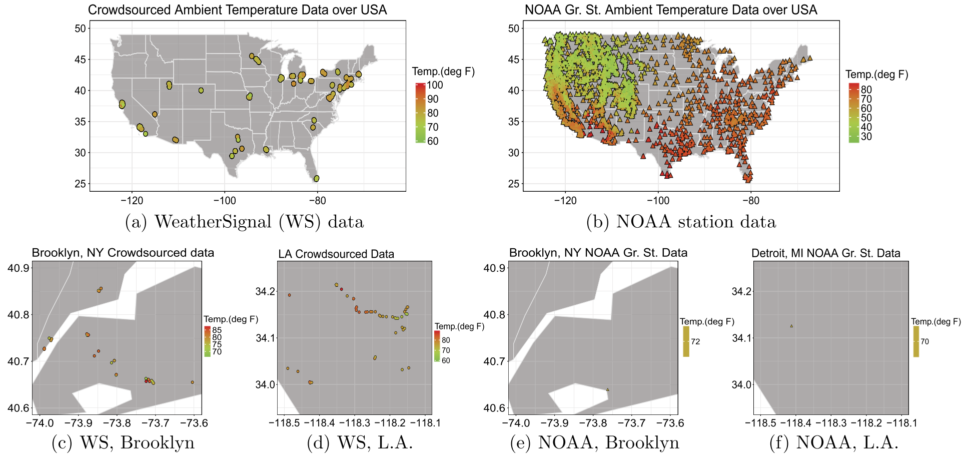
Key words and phrases. Veracity score, geostatistics, robust kriging, hyperlocal spatial prediction.

in weather information, AccuWeather, launched AccUcast in 2015 (AccuWeather (2015)), a feature that allows each user to share their local weather information as captured by the built-in mobile sensors. Other applications include Sunshine (Moynihan (2015)) and Dark Sky (Dalton (2016)), which turn each app user into a “meteorological station” for gathering and sharing hyperlocal weather information. Mobile sensor generated weather data are already being used in traffic management, fire detection, etc. In a recent article Sosko and Dalyot (2017) have used a crowdsourced mobile-sensor data in forest fire detection to densify the static geosensor network (SGN) which is primarily comprised of meteorological stations with high-performance sensors. Though spatial prediction of daily weather is generally based on satellite imagery or data from weather stations (Thornton, Running and White (1997), Vancutsem et al. (2010), Frei (2014)), recent advancement of weather-related mobile apps and the concurrent business interests call for a new methodology that considers these crowdsourced weather data to generate more accurate weather prediction in hyperlocal regions. In this article we consider the daily average ambient temperature process and show that more efficient and reasonable prediction surfaces can be created in hyperlocal regions with denser but noisy crowdsourced data as compared to a global prediction surface obtained from high-quality but coarser ground-station data.

1.1. *WeatherSignal and NOAA ground-station data.* We analyze a static crowdsourced data set consisting of geocoded daily average ambient temperature readings over the continental United States on April 30, 2013. These data were gathered by a cellphone application named `WeatherSignal`, available both for iOS and Android. In addition to providing information on current weather and forecasts, the app also gathers geographic and weather information using cellphone sensors, leading to a huge amount of crowdsourced spatial weather data from all over the globe. The `WeatherSignal` application is operated by an organization named OpenSignal. Through the research partnership program of OpenSignal, we were provided real-time (in milliseconds) ambient temperature readings captured by various mobile phones for the above-mentioned day. For each spatial location we have temporally aggregated the temperature readings to the daily average by taking mean of the regionally estimated hourly temperatures throughout the day. The details of the aggregation are explained elaborately in Section A.1 in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)). After the aggregation we have the crowdsourced daily average temperature readings at 1879 spatial locations in the United States, as shown in Figure 1(a). From the figure it can be seen that the crowdsourced observations are clumped together in high-population density regions like Detroit, Chicago, New York and Los Angeles, etc. In Figures 1(c) and 1d we show hyperlocal versions of the WeatherSignal data for two nearly square hyperlocal regions at Brooklyn, NY and Los Angeles, CA.

Along with the crowdsourced data from the WeatherSignal app, we also have ground-station data on the daily average ambient temperature from the National Oceanic and Atmospheric Administration (NOAA). We used the Global Historical Climate Network Daily (GHCND) data access tool to retrieve the daily ambient temperature summaries for April 30, 2013, from 2094 stations in the continental United States. We have plotted the ground-station observations in Figure 1(b).

Comparing Figure 1(a) and Figure 1(b), we can see that the NOAA ground-station data provides much more spatial coverage than the crowdsourced data in the entire United States or large parts of the United States, like the East Coast, Midwest, etc., are considered, and hence for global modeling or building a global prediction surface of the ambient temperature, the ground-station data is clearly a better choice. However, for hyperlocal prediction of the spatial process we believe that crowdsourced data has the potential to capture the local behavior of the spatial process more accurately. For example, in Figures 1(e) and 1(f) we have



(a) WeatherSignal (WS) data

(b) NOAA station data

(c) WS, Brooklyn

(d) WS, L.A.

(e) NOAA, Brooklyn

(f) NOAA, L.A.

FIG. 1. *Spatial plots of the crowdsourced and NOAA ground-station data.* (*c*)–(*f*) *show zoomed "hyperlocal" versions* (*each side of these regions vary from* 25 *to* 30 *miles approximately*) *of the crowdsourced WeatherSignal* (*c–d*) *and NOAA station data* (*e–f*).

plotted the available ground-station observations in the same square neighborhoods as the crowdsourced data in Figures 1(c) and 1(d). In the area around Brooklyn, NY, there are approximately 90 crowdsourced observations available, whereas the number of ground-station observations is only one. Motivated by this observation, in this paper we propose a method to improve the accuracy of the hyperlocal predictions using the available crowdsourced information in addition to the ground-station data over a bigger surrounding.

1.2. *The challenge in analyzing crowdsourced*, *mobile-sensor data.* The challenge in analyzing mobile sensor-generated crowdsourced data lies in the low quality and hence poor reliability of an unknown proportion of the data. When data are collected from mobile applications, the readings are prone to contamination for various reasons. The inaccurate observations can occur due to external factors, low-resolution sensors or a combination of these factors. For instance, the temperature readings can be affected by battery temperature, whether the user is indoor or outdoor, the proximity of the device to a hot or cold object, the heterogeneity of the sensors used by different devices and many other unknown processes.

To illustrate the varying quality of the observations in the WeatherSignal data, Figure 2 shows the temperature distribution for the two hyperlocal regions shown in Figures 1(c) and 1(d). The daily average temperature values in the crowdsourced data set vary from nearly 60°F to 90°F in both of the hyperlocal regions for the same day.

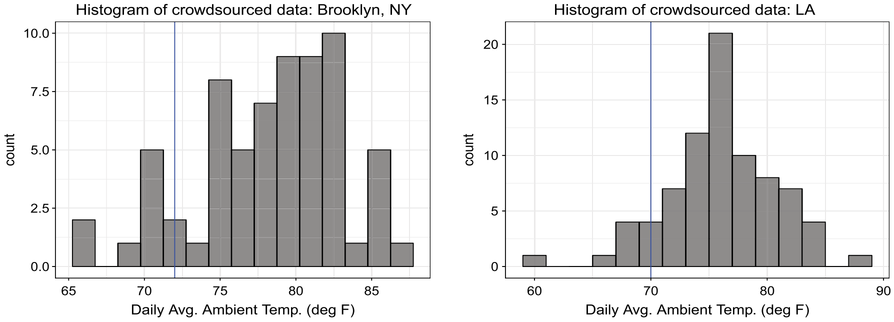


FIG. 2. *Empirical distribution of the crowdsourced average temperatures in the regions from Figure* 1 *for Brooklyn*, *NY* (*left*) *and Los Angeles*, *CA* (*right*). *Blue vertical lines represent the average ground-station values in the considered regions.*

These temperature distributions show the nature of the noise involved in the crowdsourced data. Due to the factors associated with the data collection process, a portion of the observations in the crowdsourced data are either contaminated or not representative of the ambient temperature which is the outdoor air temperature close to the earth's surface. Such representativeness errors for weather data coming from meteorological stations have been considered previously by Lorenc (1986), Gandin (1988) and Lussana, Uboldi and Salvati (2010). Comparing the histograms with the single ground-station observation in both the regions, we can see that, although there are large deviations, a good proportion of the crowdsourced observations are "close" to the corresponding ground-station observations (72°F in Brooklyn and 70°F in L.A.) which are collected in laboratory environment with high-quality sensors maintaining World Meteorological Organization (WMO) standards.

Building models based on the noisy crowdsourced data that ignore the reliability of the sensor-generated observations can lead to erroneous prediction. For instance, we used leave-one-out prediction of the observations in the regional block around Brooklyn (Figure 1(c)) using standard techniques of spatial analysis with a reasonable mean and covariance model (discussed in Section 3.1), and the errors in the predictions ranged from −30°F to 40°F. Similar cross-validation approaches have been previously used by Cressie (1993) and Lussana, Uboldi and Salvati (2010) to identify the "bad" observations. These first-stage analyses motivated us to take the quality of the observations in the WeatherSignal data into consideration. Lussana, Uboldi and Salvati (2010) proposed to remove observations for which the cross-validated prediction errors exceed some threshold. But, due to the inclusion of the corrupted observations at every iteration of the leave-one-out cross-validation, the predictions are not guaranteed to be a good representation of the true value at that location. Moreover, the leave-one-out cross-validation approach being computationally expensive, the method proposed by Lussana, Uboldi and Salvati (2010) is not readily applicable for large crowdsourced weather data coming from mobile sensors. The "absurd" observations, that is, observations with high gross error (Lussana, Uboldi and Salvati (2010)), can be identified using some other more scalable spatial outlier detection techniques (e.g., see Chapter 1 of Cressie (1993); Harris et al. (2014), etc.) and, thus, can be omitted from the analysis. But in that case it is not straightforward how to address observations with small to moderate measurement errors. For instance, using a too strict threshold on the measurement error may lead to deletion of significant number of observations, resulting in a complete loss of information for specific locations.

Hence, the new methodology should address the three following challenges. First, in addition to just identifying high-noise observations, a continuous assessment of the veracity of all the observations in a geostatistical setting is needed. Second, the definition of veracity should take into account the behavior of the process in the study region so that the "misleading" observations can be detected. Third, the veracity assessment of the observations should be incorporated into the subsequent analysis to allow for robust inference and efficient prediction. Though there are studies (e.g., Allahbakhsh et al. (2013)) in the literature on quality assessment of crowdsourced data coming from volunteers or paid participants, assessment of sensor-generated data quality is not common. Sosko and Dalyot (2017) mention an elementary root mean squared error approach for accuracy measurement using a reference data set from Israeli Meteorological Stations. However, neither of the above-mentioned papers provide full geostatistical inference and prediction using noisy crowdsourced data.

In this article we make several contributions. First, we introduce a Veracity Score (VS) to measure the reliability of the crowdsourced observations on a continuous scale using a reference data set. Second, we propose a VS-based methodology to incorporate the veracity assessment into standard spatial analysis so that the effect of noisy and misleading observations is reduced, hence making the estimation and prediction more robust and efficient. Third, we show that using the VS-based technique in hyperlocal regions with relatively higher number of crowdsourced observations can produce a more accurate and efficient prediction surface as compared to the global prediction surface obtained through the analysis of ground-station data alone. This paper is organized as follows. In Section 2 we introduce the veracity score and describe its elementary properties in a relevant geostatistical setting. Section 3 includes a brief description of the standard approach for analyzing geostatistical data, followed by a detailed description of the VS-based methodology for estimation and prediction. In Section 4 we describe simulation studies to justify the superiority of VS-based methodology over the standard approach in the analysis of noisy crowdsourced data. In Section 5 we provide details of the analysis, estimation and hyperlocal prediction in a case study. Finally, Section 6 summarizes our effort and discusses limitations and possible future works.

2. Defining and measuring veracity.

In this section we provide the intuition and motivation for veracity scoring. We denote the sample size as n. We denote the volume of a set $A \subset \mathbb{R}^2$ as $|A|$, that is, the Lebesgue measure of A if it has nonzero volume and the cardinality of A if A is finite.

2.1. *Motivation for veracity scoring.* To provide motivation for veracity scoring, consider a very simple yet practical example.

EXAMPLE 2.1. Let $Z_1, \dots, Z_n$ be independent noisy observations with $E(Z_i) = \mu$ and $\mathrm{Var}(Z_i) = \sigma_i^2$ for $i \in \{1, \dots, n\}$. The usual sample mean, which is also the o.l.s. estimator for μ, is given by $\hat{\mu}_{\mathrm{ols}} = \bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$ with $E(\hat{\mu}_{\mathrm{ols}}) = \mu$ and $\mathrm{Var}(\hat{\mu}_{\mathrm{ols}}) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2$. If we assume $\sigma_i^2 = C \cdot i^b$, for some constants (w.r.t. n) $C, b > 0$, we have

$$\mathrm{Var}(\hat{\mu}_{\mathrm{ols}}) \approx C(b) \cdot n^b$$

for some constant (w.r.t. n) $C(b)$. Instead of the generic sample mean, consider a weighted average of the observations given by $\hat{\mu} = (\sum_{i=1}^n v_i Z_i)/(\sum_{i=1}^n v_i)$, where the weights $v_i = i^{-a}$ for some constant $a > 0$, that is, the weights are inversely proportional to the variance of the noisy observations. Then,

$$\mathrm{Var}(\hat{\mu}) \approx C(a, b) \cdot n^{b-1}$$

for some constant $C(a, b)$. Clearly, if C, a and b are constants w.r.t. to the sample size n, then a significant gain in efficiency ($O(n^{b-1})$ as compared to $O(n^b)$) can be achieved for large n by assigning lower weights to high-variance observations.

If we can find a formulation of the veracity score that is inversely related to the observation noise variance, we can use it to reduce the effect of the noise in the inference and achieve a more accurate and efficient estimator.

2.2. *Preliminaries.* Let $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$ be the varying-quality observations, for example, the crowdsourced data from cell-phone sensors which are observed at irregularly spaced locations $\mathcal{S}_n := \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset \mathbb{R}^2$. In addition, at spatial locations $\mathcal{T}_m := \{\mathbf{t}_1, \dots, \mathbf{t}_m\} \subset \mathbb{R}^2$ assume that we have $\{Y(\mathbf{t}_1), \dots, Y(\mathbf{t}_m)\}$, which are high-quality, reliable observations of the spatial process, for example, measurements from the ground stations. It is common to assume (Cressie (1993), Gelfand et al. (2010)) that the spatial random field of interest $\{Y(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^2\}$ can be represented as

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{2.1}$$

where $\mu(\mathbf{s})$ is a deterministic smooth mean function capturing the large scale variation of the process, that is, $E(Y(\mathbf{s})) = \mu(\mathbf{s})$. Here, $\epsilon(\mathbf{s})$ is a mean zero spatially correlated residual process which addresses the small-scale variations over the space. For the varying-quality Z-process we write the decomposition in equation (2.1) as

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}), \tag{2.2}$$

where $w(\mathbf{s})$ is the aggregated noise associated with the observation $Z(\mathbf{s})$. For example, if we assume that the varying-quality observations arise from an additive-multiplicative noise model as

$$Z(\mathbf{s}_i) = \epsilon_{M_i} Y(\mathbf{s}_i) + \epsilon_{A_i}, \tag{2.3}$$

where ϵ_{M_i} and ϵ_{A_i} for $i \in \{1, \dots, n\}$ are random variables associated with the multiplicative and additive noise in the observation $Z(\mathbf{s}_i)$. Then, the associated w-process will have the form $w(\mathbf{s}_i) = \mu(\mathbf{s}_i)(\epsilon_{M_i} - 1) + \epsilon_{M_i}\epsilon(\mathbf{s}_i) + \epsilon_{A_i}$. If there is no multiplicative component ϵ_{M_i} in the contamination, then $w(\mathbf{s}_i) = \epsilon(\mathbf{s}_i) + \epsilon_{A_i}$. In the next subsection, we define a score to assess the quality or reliability of the observation $Z(\mathbf{s}_i)$, namely, veracity score.

2.3. *Veracity score*: *Formulation and properties*.

A good measure of veracity should not only identify “absurd” observations but also provide a score for each observation on a continuous scale, so that the effect of the “bad” observations can be reduced automatically, making inference robust against the low-quality observations. Our goal is to formulate a continuous scoring procedure to measure the veracity of the observations in two different scenarios. The first scenario assumes a reference data set containing observations with high quality but low density in the concerned regions is available. The second scenario assumes that we do not have any high-quality reference information available.

2.3.1. *Veracity score with reference data*.

Consider a hyperlocal regional block like those in Figures 1(c) or 1(d), and denote it by $\mathcal{R} \subset \mathbb{R}^2$. The observation vector with locations inside $\mathcal{R}$ is given as $\mathbf{Z} := (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$. Consider $\mathcal{R}$ to be the region of interest for analyzing the varying-quality observations $\mathbf{Z}$. Consider another regional block $\mathcal{D}$ such that $\mathcal{R} \subset \mathcal{D} \subset \mathbb{R}^2$ and $|\mathcal{R}| << |\mathcal{D}|$. Let the reference data vector with locations inside $\mathcal{D}$ be denoted as $\mathbf{Y} := (Y(\mathbf{t}_1), \dots, Y(\mathbf{t}_m))'$. The reference data $\mathbf{Y}$ is of high quality and hence reliable representation of the spatial process of interest, but it has low data coverage in the hyperlocal region of interest $\mathcal{R}$. So, to get a reasonable sample size for the reference data, we need to consider the larger region $\mathcal{D}$. We denote a δ-neighborhood around a spatial point $\mathbf{s} \in \mathbb{R}^2$ as $\mathcal{B}_\delta(\mathbf{s})$ with $\mathcal{B}_\delta(\mathbf{s}) := (\mathbf{s} - \delta, \mathbf{s} + \delta]$ for some $\delta \in \mathbb{R}^+$ where the subtraction and addition is componentwise.

Define the VS of the observation $Z(\mathbf{s}_i)$ as

$$V(\mathbf{s}_i) = \phi\left(\frac{|Z(\mathbf{s}_i) - \xi(\mathbf{s}_i)|}{\alpha + D(\boldsymbol{\xi}_i)}\right), \tag{2.4}$$

where $\phi : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ is some nonincreasing function such that $\sup_x \phi(x) < \infty$. We call $\phi(\cdot)$ the veracity function with $\alpha \in \mathbb{R}^+$ as a regularity parameter. By $\xi(\mathbf{s}_i)$, we denote a reasonable benchmark for the target process at $\mathbf{s}_i$, and $\boldsymbol{\xi}_i := (\xi(\mathbf{s}_{i_1}), \dots, \xi(\mathbf{s}_{i_{n(i)}}))'$ where $\{\mathbf{s}_{i_1}, \dots, \mathbf{s}_{i_{n(i)}}\}$ is the set of observation locations in the small δ-neighborhood $\mathcal{B}_\delta(\mathbf{s}_i)$. Finally, $D(\mathbf{x})$ denotes a robust measure of dispersion of the observations in the vector $\mathbf{x}$. Clearly, the VS is computed by evaluating the ϕ-function at the *scaled deviation* $\frac{|Z(\mathbf{s}_i)-\xi(\mathbf{s}_i)|}{\alpha+D(\boldsymbol{\xi}_i)}$ and due the nonincreasing property of $\phi(\cdot)$; if the deviation is high, we have low VS, and if the deviation is low, we have high VS.

Now consider the benchmark value, $\xi(\mathbf{s})$, for the target at location $\mathbf{s}$. If we have high-quality observations of the Y-process from the reference data at the varying-quality data sites $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, then the obvious choice is to take $\xi(\mathbf{s}_i) = Y(\mathbf{s}_i)$. In practice, as we see in Figures 1(c) to 1(f), the locations of the ground-station measurements (reference data) and the crowdsourced data (varying-quality observations) almost always differ significantly. Hence, to define the benchmark at location $\mathbf{s}_i$, we propose to compute a kriging surface, $\{\mathbf{s}, \hat{Y}(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, of the Y-process using the observation vector $\mathbf{Y}$. Then, we define $\xi(\mathbf{s}_i)$ as

$$\xi(\mathbf{s}_i) = \hat{Y}(\mathbf{s}_i) + (1 - \nu)\mathcal{C}(\mathbf{Z}_i - \hat{\mathbf{Y}}_i), \tag{2.5}$$

where $\mathbf{Z}_i := (Z(\mathbf{s}_{i_1}), \dots, Z(\mathbf{s}_{i_{n(i)}}))'$ and $\hat{\mathbf{Y}}_i := (\hat{Y}(\mathbf{s}_{i_1}), \dots, \hat{Y}(\mathbf{s}_{i_{n(i)}}))'$. Here, $\mathcal{C}(\mathbf{x})$ is a robust measure of central tendency of the values in the vector $\mathbf{x}$, and $\nu \in [0, 1]$ is a mixing parameter that we discuss in detail later.

If we have a reasonable benchmark, $\xi(\mathbf{s}_i)$, for the spatial process of interest at the location $\mathbf{s}_i$ the definition of the VS in equation (2.4) is a transformed measure of the scaled deviation of the observation $Z(\mathbf{s}_i)$ from the benchmark value. In the definition of VS, the measure of dispersion, $D(\boldsymbol{\xi}_i)$, in the denominator takes the variability in the δ-neighborhood into account. For example, in the analysis of ambient temperature the variation in a small neighborhood in the mountains is likely to be higher than an area close to sea level. Hence, the statistic $\frac{|Z(\mathbf{s}_i) - \xi(\mathbf{s}_i)|}{\alpha + D(\boldsymbol{\xi}_i)}$ measures the deviation of the observation from its benchmark relative to the local variability. In the following sections we use interquartile range (i.e., $D(\mathbf{x}) = \mathrm{IQR}(\mathbf{x})$) as the robust measure of dispersion in equation (2.4) and the sample median (i.e., $\mathcal{C}(\mathbf{x}) = Q_2(\mathbf{x})$ where Q_j is the jth sample quartile) as the robust measure of central tendency in equation (2.5). There are other robust choices as well, but we use the sample quantile based statistic because it is familiar to the practitioners and easy to interpret. Also, these choices are theoretically justified as the sample quantiles are asymptotically consistent under dependence (Ghosh (1971), Sun and Lahiri (2006)). The parameter α determines the baseline of the deviation. For lower values of α we penalize more, and for higher values we allow for a larger deviation from the benchmark. We call α the *baseline deviation* of the VS, and its unit is same as the process of interest which makes the VS unit free.

We require the veracity function ϕ to have the following properties:

1. $\phi(\cdot)$ is a nonincreasing function with bounded range, $\phi(x) \le \phi(0) < \infty$.
2. $\phi(x) \downarrow 0$ as $x \to \infty$.

With this formulation, lower values of the VS correspond to the low-quality or less-reliable observations and high values of the VS correspond to the better quality of the observations. We use $\phi(x) = \exp(-x)$ for our analysis in the subsequent sections. The advantage of this function is that the VS lies naturally in [0, 1], and it penalizes exponentially as the scaled deviation from the benchmark value increases. We discuss other possible choices in Section B.1 in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)).

Now we try to interpret the mixing parameter ν in the definition of VS. Under the assumption that the estimated mean process $\hat{\mu}(\mathbf{s})$ is smooth and the kriged-residual process $\hat{\epsilon}(\mathbf{s})$ is a spatially correlated second-order stationary mean-zero process, for a small enough $\delta > 0$, we can write $Q_2(\hat{\mathbf{Y}}_i) \approx \hat{Y}(\mathbf{s}_i)$ as the variation of the kriged process $\hat{Y}(\mathbf{s})$ inside the δ-neighborhood is negligible. Hence, we can approximately rewrite the benchmark as

$$\xi(\mathbf{s}_i) \approx \nu\hat{Y}(\mathbf{s}_i) + (1 - \nu)Q_2(\mathbf{Z}_i).$$

Here, to get a possible approximation the spatial process at location $\mathbf{s}_i$, instead of just using the estimated value $\hat{Y}(\mathbf{s}_i)$ from the high-quality reference data over a bigger surrounding, we want to leverage the available varying-quality observations in the hyperlocal region. We propose to use a mixture of an approximation of the spatial process coming from the reference data over a bigger region $\mathcal{D}$, that is, $\hat{Y}(\mathbf{s}_i)$ and a robust local estimate coming from the varying-quality observations in the small δ-neighborhood $\mathcal{B}_\delta(\mathbf{s}_i)$ around the location of interest $\mathbf{s}_i$, that is, $Q_2(\mathbf{Z}_i)$. Due to the smooth mean and spatially correlated residual process, the spatial observations in a "small" neighborhood are likely to behave "similarly." Therefore, it is sensible to use a robust estimate of the central tendency of the varying-quality observations in that small neighborhood as the locally estimated approximation of the spatial process at $\mathbf{s}_i$. The mixing parameter ν decides the weight of mixing between the estimated process from the reference data and the local approximation from the varying-quality observations. The optimal ν balances the error in estimation from the reference data and the error in the approximation of the spatial process using the sample median in the δ-neighborhood.

2.3.2. *Veracity score without reference data.* We propose a similar definition of the VS when we do not have any high-quality reference observations available. In this scenario our definition of VS is

$$V(\mathbf{s}_i) = \phi\left(\frac{|Z(\mathbf{s}_i) - \mathcal{C}(\mathbf{Z}_i)|}{\alpha + D(\mathbf{Z}_i)}\right). \tag{2.6}$$

The idea behind the definition given in equation (2.6) is similar to that in Section 2.3.1. As we do not have information available from a high-quality reference data set, we use only the locally estimated central tendency as the proxy of the target and the local variation in the denominator to take the regional variability into account. Note that the definition of the VS in equation (2.4) approximately equals the VS as given in equation (2.6) if we take $\nu = 0$.

The formulations of the VS, both with and without reference data, depend on δ which is a positive scalar equal to half of the length of the neighborhood $\mathcal{B}_\delta(\mathbf{s}_i)$ used to estimate the center and dispersion locally. The choice of δ should be such that the δ-neighborhood $\mathcal{B}_\delta(\mathbf{s}_i)$ is small as compared to the region of interest $\mathcal{R}$ but at the same time large enough to have sufficient sample size to provide a good assessment of the quality of the observations. To make the formulation of VS as well defined, we need the number of points in the δ-neighborhood, $n(i)$, larger than two for each $i \in \{1, 2, \dots, n\}$. If we do not have enough data points to compute the measure of dispersion for an observation, we say that the VS is undefined for those observations.

A similar approach of comparing the observations with a *benchmark* value has been used to detect outliers in literature (e.g., see Chapter 1 of Cressie (1993); Papritz (2018a)). Lussana, Uboldi and Salvati (2010) proposed a benchmark obtained through leave-one-out cross-validated prediction using the noisy observations. But, as mentioned in Section 1.2, due the presence of some absurd noise in the training data of the cross-validation, the benchmarks obtained in this technique might themselves be corrupted and hence, are not necessarily robust. We prefer quantile based local summaries as benchmarks due to its scalability and computational ease, appeal to the practitioners as well as robustness and asymptotic efficiency (see Sen (1968)) as compared to some other choices discussed previously.

3. Veracity score methods.

Before going to the VS-based version of the spatial analysis, we briefly describe the standard approach of geostatistical analyses.

3.1. *Review of standard analysis of spatial data.* For this section we use the model specified in equations (2.1) and 2.2 as well as the notations stated in Section 2.2. In geostatistics often the smooth deterministic mean process $\{\mu(\cdot)\}$ is modeled under a spatial regression framework where the mean function is assumed to have a *linear* form, $\mu(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$, where

$\mathbf{x}(\cdot) = (x_1(\cdot), \dots, x_p(\cdot))'$ is a p-dimensional deterministic vector process of known covariates and $\boldsymbol{\beta}$ denotes the unknown regression parameter vector. To make the inference feasible from only one replication of the process over the space, some stationarity assumption on the second-order structure of the residual process $\{\epsilon(\mathbf{s})\}$ is required. One of the most commonly used assumptions is that $\{\epsilon(\mathbf{s})\}$ is an intrinsically stationary process with an admissible parametric variogram function $2\gamma(\mathbf{h}; \boldsymbol{\theta}) = \mathrm{Var}\{\epsilon(\mathbf{s}) - \epsilon(\mathbf{s}+\mathbf{h})\}$ where $\boldsymbol{\theta}$ is the covariance parameter of interest.

For now, the description of the analysis is given without taking the noisy nature of the observations into account, so $\{w(\mathbf{s})\}$ is assumed to be identically equal to $\{\epsilon(\mathbf{s})\}$. Since the covariance parameter is unknown, the standard analysis starts with the estimation of the regression parameters in the linear mean model using ordinary least squares (o.l.s.), $\hat{\boldsymbol{\beta}}_{\mathrm{ols}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ where, $\mathbf{X} := (\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n))'$. Next, the detrended observations, that is, $\hat{\boldsymbol{\epsilon}} = \mathbf{Z} - X\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$, are used to estimate the covariance parameter $\boldsymbol{\theta}$ using least squares-based variogram model fitting (Cressie (1993)) based on some generic nonparametric semivariogram estimator (denoted by $\hat{\gamma}(\mathbf{h})$), for example, the *classical* or *method-of-moments* semivariogram estimator proposed by Matheron (1962). For example, the weighted least squares (w.l.s.) estimator of $\boldsymbol{\theta}$ is given as

$$\hat{\boldsymbol{\theta}}_{\mathrm{wls}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{j=1}^{k} w_j \{\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j; \boldsymbol{\theta})\}^2, \tag{3.1}$$

where w_j is the weight corresponding to lag $\mathbf{h}_j$ and, $\{\mathbf{h}_1, \dots, \mathbf{h}_k\}$ are the set of discrete lags for which the nonparametric semivariogram $\hat{\gamma}(\cdot)$ has been computed. For details of variogram model fitting, see Cressie (1993), Gelfand et al. (2010). Matérn is a popular choice for the parametric class of admissible variograms as it provides a rich class to choose from (Haskard (2007)). A comprehensive list of parametric variogram models can be found in Cressie (1993) and Gneiting (2013).

Once the covariance structure is estimated, one can try to improve the mean parameter estimates using *estimated generalized least squares* (e.g.l.s.) estimator, given by $\hat{\boldsymbol{\beta}}_{\mathrm{egls}} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}\mathbf{Z}$ where $\hat{\Sigma}$ is the estimated variance of $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \dots, \epsilon(\mathbf{s}_n))'$. However, this introduces additional variability due to using the estimated covariance parameters in the mean estimator and is not necessarily more efficient than the o.l.s. estimator.

The most commonly used method to predict the process at new locations is to predict the ϵ-process at the given locations by the *best linear unbiased predictor* (BLUP) given the observed residual vector $\hat{\boldsymbol{\epsilon}}$, also known as *ordinary kriging* estimator (Cressie (1993), page 122). The standard predictor of $Y(\mathbf{s}_0)$ is

$$\hat{Y}_{\mathrm{std}}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\hat{\boldsymbol{\beta}}_{\mathrm{ols}} + \hat{\epsilon}_{\mathrm{ok}}(\mathbf{s}_0), \tag{3.2}$$

where $\hat{\epsilon}_{\mathrm{ok}}(\mathbf{s}_0)$ is the ordinary kriging predictor for $\epsilon(\mathbf{s}_0)$.

The standard approach for estimation and prediction explained is not reliable for analyzing noisy spatial observations, as both the least squares-based mean parameter estimators (Huber and Ronchetti (2009)) and the method-of-moments empirical semivariogram estimator are highly sensitive to the noise (Cressie and Hawkins (1980)) in the data. In the following sections we propose a way to incorporate the VS into the analysis to make the inference and prediction robust against the noise in the data.

3.2. *Veracity score-based estimation of the mean function.* In the standard approach, as described in Section 3.1, the regression parameter vector $\boldsymbol{\beta}$ is estimated using the o.l.s. method. For our approach, instead of simple squared error loss, motivated by Example 2.1, we propose to minimize a weighted version of the loss function with the veracity scores as the corresponding weights. The VS-based estimator of the mean parameter $\boldsymbol{\beta}$ is given as

$$\hat{\boldsymbol{\beta}}_{\mathrm{VS}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}\big(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta}\big). \tag{3.3}$$

For least squares-based estimators we have $\mathcal{L}(y, u) = (y - u)^2$, the squared-error loss function. The locally estimated veracity scores lessen the effects of "absurd" observations in the objective function and thus make the estimation of the mean function less sensitive to the noise. The VS-based approach is adaptive to the quality of the observations and thus lessens the impact of outliers in the data. To make the estimation more robust to contamination, one can use any robust loss function instead of squared-error loss in equation (3.3). We have used an MM-type estimator with a *linear quadratic quadratic* ψ-function for the robust regression as discussed in Koller and Stahel (2011). The advantage of using this estimator is that, in addition to penalizing less for high residuals, the parameters associated with the ψ-function can be tuned to improve the asymptotic efficiency for the estimators. The corresponding optimization to solve equation (3.3) can be executed using Iterative Reweighted Least Squares (IRLS) as discussed in Todorov and Filzmoser (2009).

The assessment of goodness of fit for the estimated linear model is essential. The usual Multiple R^2 is not reasonable to use, as the loss function is different from ordinary least squares. Inspired by the pseudo-$\mathrm{R}^2_{\mathrm{WLS}}$ coined by Willet and Singer (1988), we propose another variant of the coefficient of determination for VS-based regression as

$$\mathrm{R}^2_{\mathrm{VS}} = 1 - \frac{\sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}_{\mathrm{VS}})}{\sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}(Z(\mathbf{s}_i), \bar{Z})},$$

where $\bar{Z} = n^{-1}\sum_i Z(\mathbf{s}_i)$. The idea behind this measure is that instead of using the squared error loss to compute the total sum of squares and the residual sum of squares, the proposed $\mathrm{R}^2_{\mathrm{VS}}$ uses the robust loss function to measure the total variability in the data (i.e., $\sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}(Z(\mathbf{s}_i), \bar{Z})$) and the variability that is not explained by the model (i.e., $\sum_{i=1}^{n} V(\mathbf{s}_i)\mathcal{L}(Z(\mathbf{s}_i), \mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}_{\mathrm{VS}})$). Although we do not provide any theoretical justification, it appears from explanatory analysis with synthetic data and simulations that $\mathrm{R}^2_{\mathrm{VS}}$ may provide an overly optimistic assessment of the goodness of the fit for the model when the Huber's loss function or MM-type estimation is used.

3.3. *Veracity score-based estimation of the covariance structure.* To explore the second-order structure of the spatial process, we analyze the residuals obtained by detrending the observations $\hat{\epsilon}_{\mathrm{VS}}(\mathbf{s}_i) = Z(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}_{\mathrm{VS}}$ for $i \in \{1, 2, \dots, n\}$. When conducting analysis with varying quality geostatistical data, after the robust estimation of the regression parameters a portion of the residuals are affected by the presence of measurement error in the data, and direct analysis of these residuals can result in misleading and inefficient estimation of the covariance structure. To reduce the noise of the observed residuals, we propose a VS-based modification of residuals using a local smoothing prior estimation of the covariance parameters. When we have a high-quality reference data, we define the VS-based smoothed version of the residuals as

$$\tilde{\epsilon}(\mathbf{s}_i) = V(\mathbf{s}_i)^q \hat{\epsilon}_{\mathrm{VS}}(\mathbf{s}_i) + \big(1 - V(\mathbf{s}_i)^q\big) Q_2(\boldsymbol{\xi}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\mathrm{VS}}), \tag{3.4}$$

where $\mathbf{X}_i := (\mathbf{x}(\mathbf{s}_{i_1}), \dots, \mathbf{x}(\mathbf{s}_{i_{n(i)}}))'$ is the $n(i) \times p$ matrix of the covariates corresponding to the observations in $\mathcal{B}_\delta(\mathbf{s}_i)$. Here, q is the parameter regulating the degree of the smoothing needed. For instance, $q = 0$ implies no smoothing, and $q = 1$ implies the convex combination of the locally-corrected residual and the observed residual. As shown in Figure S4-(a) in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)), the parameter q here plays the role of thresholding for higher q, only observed residuals with high VS get significant weights for the VS-based smoothing, whereas for smaller q the formulation of the smoothed residuals in equation (3.4) puts significant weights to even the observed residuals with low VS and thus reducing the degree of smoothing.

If we do not have reference data available, then the analogous smoothed version of the residuals is given by

$$\tilde{\epsilon}(\mathbf{s}_i) = V(\mathbf{s}_i)^q \hat{\epsilon}_{\mathrm{VS}}(\mathbf{s}_i) + \big(1 - V(\mathbf{s}_i)^q\big) Q_2(\hat{\boldsymbol{\epsilon}}_i), \tag{3.5}$$

where $\hat{\boldsymbol{\epsilon}}_i = (\hat{\epsilon}_{\mathrm{VS}}(\mathbf{s}_{i_1}), \dots, \hat{\epsilon}_{\mathrm{VS}}(\mathbf{s}_{i_{n(i)}}))'$. Again, note that the definition in equation (3.4) approximately simplifies to the one in equation (3.5) if $\nu = 0$.

For poor quality observations, when $V(\mathbf{s}_i)$ is small, the effect of the observed value of the residual $\hat{\boldsymbol{\epsilon}}_{\mathrm{VS}}(\mathbf{s}_i)$ is scaled down by $V(\mathbf{s}_i)^q$ (as $V(\mathbf{s}_i) \in (0, 1]$), and the locally estimated "benchmark" value of the residual process in the small neighborhood is enforced by $(1 - V(\mathbf{s}_i)^q)$ in equations (3.4) and 3.5. The effect of VS-based smoothing is illustrated on a synthetic data set in Section B.2 and Figure S.3 in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)).

We propose to use variogram model fitting with the VS-based smoothed version of the residuals, $\{\tilde{\epsilon}(\mathbf{s}_i)\}_{i=1}^n$ to estimate the covariance parameter $\boldsymbol{\theta}$ robustly. First, a generic nonparametric semivariogram is evaluated at discrete lags using the *robust* semivariogram estimator proposed by Cressie and Hawkins (1980),

$$\hat{\gamma}_{\mathrm{VS}}(\mathbf{h}_u) = \frac{\{\frac{1}{2|N(H_u)|}\sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(H_u)}|\tilde{\epsilon}(\mathbf{s}_i) - \tilde{\epsilon}(\mathbf{s}_j)|^{\frac{1}{2}}\}^4}{0.457 + \frac{0.494}{|N(H_u)|}} \quad \text{for } u \in \{1, \dots, K\}, \tag{3.6}$$

where $N(H_u) = \{\mathbf{h} \in \mathcal{H} : \mathbf{h} \in H_u\}$. H_u are small lag classes or *bins* (see page 34, Gelfand et al. (2010)), which are often called *tolerance regions* (see page 70, Cressie (1993)), and these construct a partition of size K of the lag space $\mathcal{H} = \{\mathbf{s} - \mathbf{s}' : \mathbf{s}, \mathbf{s}' \in \mathcal{R}\}$. The candidate lag for the tolerance region H_u is denoted by $\mathbf{h}_u$ which is often taken to be the mean of the observed lags in the bin or the centroid of the the class H_u.

The parameters are estimated using method of weighted least squares as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{VS}} &= \operatorname*{argmin}_{\boldsymbol{\theta}} Q_{\mathrm{wls}}(\boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{u=1}^{K} \frac{|N(\mathbf{h}_u)|}{\{\gamma(\mathbf{h}_u; \boldsymbol{\theta})\}^2} \{\hat{\gamma}_{\mathrm{VS}}(\mathbf{h}_u) - \gamma(\mathbf{h}_u; \boldsymbol{\theta})\}^2, \end{aligned} \tag{3.7}$$

where $\gamma(\cdot; \boldsymbol{\theta})$ is some prespecified parametric admissible semivariogram model, as discussed in Section 3.1. Other robust empirical variogram estimators (e.g., Genton (1998), Lark (2000)) can also be used instead of the one proposed by Cressie and Hawkins (1980), as given in equation (3.6), Genton (1998) showed that the robustness properties of the empirical semivariogram proposed by Cressie and Hawkins (1980) are not enough in the presence of "absurd" outliers in the data. But, due to the VS-based smoothing in the first stage of the covariance estimation, the very large measurement errors have already been addressed and, hence, using Cressie and Hawkins (1980) version of robust variogram estimator is reasonable here.

3.4. *Veracity score-based spatial prediction.* Often the aim for spatial analysis of geostatistical data is to predict the process at locations of interest or to create a prediction surface over a region of interest. To predict the ϵ-process at a new location $\mathbf{s}_0$, we can use ordinary kriging with the VS-based smoothed residuals $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}(\mathbf{s}_1), \dots, \tilde{\epsilon}(\mathbf{s}_n))'$ as

$$\tilde{\epsilon}(\mathbf{s}_0) = \left\{\boldsymbol{\gamma} + \mathbf{1}\frac{(1 - \mathbf{1}'\Gamma^{-1}\gamma)}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}\right\}' \Gamma^{-1}\tilde{\boldsymbol{\epsilon}}, \tag{3.8}$$

where $\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0 - \mathbf{s}_1; \hat{\boldsymbol{\theta}}_{\text{vs}}), \dots, \gamma(\mathbf{s}_0 - \mathbf{s}_n; \hat{\boldsymbol{\theta}}_{\text{vs}}))'$ and $(\Gamma)_{ij} = \gamma(\mathbf{s}_i - \mathbf{s}_j; \hat{\boldsymbol{\theta}}_{\text{vs}})$ (see Chapter 3, Cressie (1993)). The residual kriging variance, which quantifies the prediction uncertainty, can be estimated as

$$\hat{\text{Var}}(\tilde{\epsilon}(\mathbf{s}_0)) = \hat{\sigma}^2_{\text{ok}}(\mathbf{s}_0) = \boldsymbol{\gamma}'\Gamma^{-1}\boldsymbol{\gamma} - \frac{(\mathbf{1}'\Gamma^{-1}\gamma)^2}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}.$$

Finally, we predict the process at $\mathbf{s}_0$ using the modified version of equation (3.2) as

$$\hat{Y}_{\text{vs}}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\hat{\boldsymbol{\beta}}_{\text{vs}} + \tilde{\epsilon}(\mathbf{s}_0). \tag{3.9}$$

In equation (3.9) both the mean and covariance parameters have been robustly estimated using the VS-based procedures. The smoothing parameter q for the VS-based smoothing of the residuals can be chosen using cross-validation.

There are other robust kriging approaches available in literature, for example, Künsch et al. (2011) and Papritz (2018b). Both of these techniques require distributional assumption on the ϵ-process. Moreover, it is not straightforward to determine how to reduce the effects of observations that are not noisy but represent some other spatial process. For example, if in a local region most of the crowdsourced ambient temperatures are captured in indoor settings, applying the robust procedures directly may lead to misleading estimation of the model parameters and hence bad prediction of the outdoor ambient temperature. On the other hand, the VS-based technique can use a benchmark value, possibly obtained from a high-quality but low-density reference data, to reduce the effects of the "misleading" observations and thus estimate and predict the process of interest efficiently. Theoretical or numerical comparison of other robust kriging methodologies with the VS-based technique in case of no available reference data is beyond the scope of this article.

4. Simulation study.

Our simulation study aims to justify the superiority of the VS-based estimation and prediction methods as compared to the standard approach for analyzing noisy geostatistical data. We have considered two scenarios here. The first one is when no reference data is available, and the second one is when a coarser but better quality reference data is present.

4.1. *Without reference data.* We take the sampling region for the varying-quality observations to be $\mathcal{R} \equiv \mathcal{R}_n := [0, \lambda_n]^2$ where $\{\lambda_n\}_n$ is a sequence of positive real numbers determining the size of the sampling region. We have assumed that the varying-quality observations $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$ are coming from an additive-multiplicative noise model as given in equation (2.3). To generate the "true" process for simulation purposes, we use the following spatial linear model:

$$Y(\mathbf{s}_i) = \beta_0 + (\beta_x, \beta_y)'\mathbf{s}_i + \beta_h h(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \tag{4.1}$$

where $\boldsymbol{\beta} := (\beta_0, \beta_x, \beta_y, \beta_h)'$ is the vector of regression parameters; $h(\mathbf{s})$ is the altitude of the location $\mathbf{s}$; and $\{\epsilon(\mathbf{s})\}$ is a second-order stationary spatially correlated process.

To define the altitude function over the sampling region, we use the deterministic function $h(\mathbf{s}) = H_1 \cdot \sum_{j=1}^{H_2} w_h(j) f(\mathbf{s}; \boldsymbol{\mu}_j, \Sigma_j) + H_3$ where $f(\cdot; \boldsymbol{\mu}, \Sigma)$ denotes the bivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix Σ and $\{(\boldsymbol{\mu}_j, \Sigma_j) : j \in \{1, \dots, H_2\}\}$ are fixed set of vectors and matrices. The residual vector $(\epsilon(\mathbf{s}_1), \dots, \epsilon(\mathbf{s}_n))'$ are sampled from a second-order stationary mean-zero Gaussian process with isotropic Matérn covariance given by

$$C(d; \boldsymbol{\theta}) = \sigma_\epsilon^2 \frac{2^{1-\kappa}}{\Gamma(\kappa)} \left(\sqrt{2\kappa}\frac{d}{\rho}\right)^\kappa K_\kappa\left(\sqrt{2\kappa}\frac{d}{\rho}\right) + \tau^2 \mathbb{1}(d = 0), \tag{4.2}$$

where Γ is the gamma function and K_κ is the modified Bessel function of the second kind with order κ (Abramowitz and Stegun (1972)). The covariance parameter vector of interest is $\boldsymbol{\theta} = (\tau^2, \sigma_\epsilon^2, \rho, \kappa)'$, where τ^2 is the nugget effect, σ_ϵ^2, ρ, κ are the partial sill, range and smoothness parameters, respectively (Haskard (2007), Gelfand et al. (2010)).

To generate noise for the varying-quality observations, we use the following model for the additive and multiplicative components, denoted by $\boldsymbol{\epsilon}_A := (\epsilon_{A_1}, \dots, \epsilon_{A_n})'$ and $\boldsymbol{\epsilon}_M := (\epsilon_{M_1}, \dots, \epsilon_{M_n})'$, respectively

$$\epsilon_{M_i} \sim \begin{cases} \Delta(1) & \text{if } i \in G_n, \\ 2 \times \text{Beta}(\alpha_M, \alpha_M) & \text{o.w.,} \end{cases} \qquad \epsilon_{A_i} \sim \begin{cases} \Delta(0) & \text{if } i \in G_n, \\ N(0, \sigma_A^2) & \text{o.w.,} \end{cases} \tag{4.3}$$

where $\Delta(x)$ denotes a degenerate distribution with point mass at $-\infty < x < \infty$; variance corresponding to the multiplicative component $\sigma_M^2 = \frac{1}{2\alpha_M+1}$; $G_n \subset \{1, \dots, n\}$ is a subset of indices and σ_M, σ_A are positive constants. With this model, if $i \in G_n$, we have no noise associated with the observation, that is, $Z(\mathbf{s}_i) = Y(\mathbf{s}_i)$. If $i \notin G_n$, then $Z(\mathbf{s}_i) = \epsilon_{M_i} Y(\mathbf{s}_i) + \epsilon_{A_i}$ where ϵ_{M_i} and ϵ_{A_i} have positive variance. Also, we have taken $\{\epsilon_{M_i}\}_{i=1}^n$, $\{\epsilon_{A_i}\}_{i=1}^n$ and $\{\epsilon(\mathbf{s}_i)\}_{i=1}^n$ are independent of each other. We further assume that the proportion of "good" observations is a constant (w.r.t. n) denote by q_e, that is, $|G_n|/n \approx q_e$, and $1 - q_e$ is the proportion of noisy observations in the data. This model is inspired by the crowdsourced data analysis scenario where only a proportion of observations are "bad." The choice of multiplicative error distribution in equation (4.3) restricts its realizations to be in $[0, 2]$ and also ensures that the multiplicative errors are symmetric around 1.

We set $\boldsymbol{\beta} = (55, 1.5, -1, -0.08)'$, $\boldsymbol{\theta} = (0, 6, 0.5, 3)'$. To investigate the robustness of the VS with increasing noise in the data, we consider three contamination models specified by the following parameters: (a) $\sigma_A = 5$, $\alpha_M = 2$, $q_e = 0.95$, (b) $\sigma_A = 50$, $\alpha_M = 0.5$, $q_e = 0.9$ and (c) $\sigma_A = 100$, $\alpha_M = 0.05$, $q_e = 0.8$. As we go from model (a) to (c), the noise in the data increases both in extent and magnitude. For example, with model (a) the variance of a noisy observation at location $\mathbf{s}$ is $0.2(\mathbf{x}(\mathbf{s})'\boldsymbol{\beta})^2 + 28.6$, and the proportion of such observations is 5%; with model (c) the same variance will be $0.91(\mathbf{x}(\mathbf{s})'\boldsymbol{\beta})^2 + 10{,}005.73$, and the proportion of noisy observations rises to 20%.

Next, we analyze the simulation results to compare the performances of VS-based and standard approach. The choices of the regularity parameters in the VS-based estimation like the baseline deviation α and the smoothing parameter q are discussed in Section C.1 in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)).

In Figure 3 we show boxplots of the VS-based estimator $\hat{\beta}_{\text{vs}}$ and the standard estimator $\hat{\beta}_{\text{ols}}$ for the four regression parameters based on $B = 200$ simulations with $n = 500$ samples. The VS-based technique shows more robustness toward the added noise in the observations. As we move from noise model (a) to (c), the efficiency of the o.l.s. estimator is heavily compromised, whereas the spread of the VS-based estimates is hardly increased. Section C.2 in the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)) contains additional simulation results for regression parameter estimation; boxplots if the estimates for $n = 100, 3000$ (Figures S4 and S5). All of these simulations show similar results to justify the superiority of

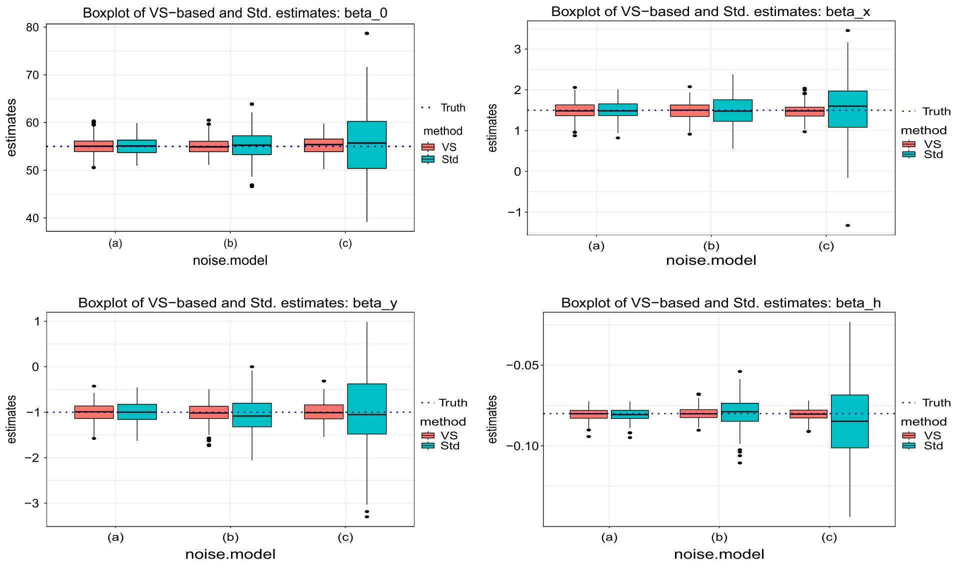


FIG. 3. *Performance of the VS-based and standard regression parameter estimators for analyzing varying-quality observations* (*sample size* $n = 500$) *without reference data.*

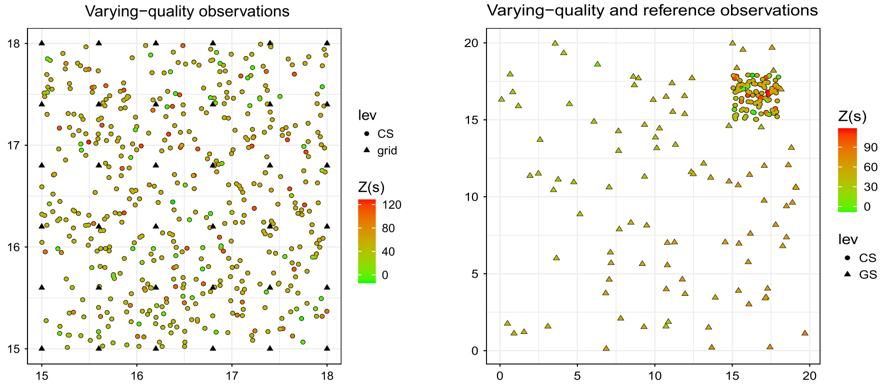
VS-based mean parameter estimation in the analysis of noisy spatial data as compared to the standard o.l.s method.

We also evaluate the VS-based and standard covariance parameter estimation and show the results in Table 1. In each of the cases, the estimates of the sill parameter ($\sigma_\epsilon^2 + \tau^2$, the total variance the residual process) obtained by the VS-based methodology is more accurate by large margins as compared to standard variogram estimation. As the sample size increases both the bias and standard deviations of the VS-based estimators are closing toward *zero* under all the considered noise models. Table 1 clearly establishes the efficiency of VS-based covariance estimation as compared to the standard approach when some of the observations are corrupted. For a fixed n, if we move from noise model (a) to noise model (c) the increase in bias and standard errors of the VS-based sill parameter estimator is prominent, though the magnitude of increment is much smaller as compared to the standard method of estimation.

Next, we evaluate the VS-based spatial prediction using a $4\lceil\lambda_n\rceil \times 4\lceil\lambda_n\rceil$ grid over the sampling region $\mathcal{R}$ as shown in Figure 4(a). We make predictions at these grid points using both the VS-based and standard approach and evaluate the predictions and kriging by the

TABLE 1
Performance of the VS-based methodology and standard approach in estimating covariance parameters on varying-quality observations

Noise model	n	bias.sill.VS	bias.sill.Std	bias.range.VS	bias.range.Std
(a)	100	−0.313 (3.31)	3837.513 (9867.28)	−0.296 (0.13)	6.671 (16.1)
	500	0.23 (1.16)	623.629 (1644.56)	−0.114 (0.06)	3.778 (9.91)
	3000	0.344 (0.62)	36.098 (82.01)	−0.026 (0.05)	0.307 (3.2)
(b)	100	7.657 (8.13)	17,545.465 (58,680)	−0.357 (0.08)	69.945 (454.78)
	500	1.747 (1.52)	5135.181 (14,207.51)	−0.158 (0.06)	3.711 (11.05)
	3000	0.48 (0.96)	1108.544 (3515.95)	−0.06 (0.05)	8.377 (52.63)
(c)	100	32.774 (9.51)	6606.713 (27,599.4)	−0.39 (0.03)	130.833 (463.05)
	500	15.352 (6.23)	21,915.533 (63,507.44)	−0.241 (0.05)	6.222 (47.09)
	3000	2.933 (1.14)	5289.832 (12,192.3)	−0.111 (0.04)	4.624 (19.92)



(a) Varying-quality observations (CS = crowdsourced) and the grid to validate prediction.

(b) Varying-quality hyperlocal observations (CS) with reference data (GS = ground-station).

FIG. 4. *Example sampling points for the simulations.*

following metrics:

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{\mathbf{s}^*}\big(\hat{Y}_{\mathrm{VS}}(\mathbf{s}^*) - Y(\mathbf{s}^*)\big)^2};$$

$$\mathrm{ResRMSPE} = \sqrt{\frac{1}{n}\sum_{\mathbf{s}^*}\big(\tilde{\epsilon}(\mathbf{s}^*) - \epsilon(\mathbf{s}^*)\big)^2},$$

where the sum $\sum_{\mathbf{s}^*}$ is over the grid points. We define the performance metrics for the standard methods analogously. The Root-Mean-Squared-Prediction-Error (RMSPE) measures the average prediction error over the selected grid, and the Residual-Root-Mean-Squared-Prediction-Error (ResRMSPE) evaluates the accuracy and efficiency of the kriging on the selected grid for the spatially correlated residual process, $\{\epsilon(\mathbf{s})\}$. By Av.RMSPE we denote $\frac{1}{B}\sum_b \mathrm{RMSPE}(b)$ where $\mathrm{RMSPE}(b)$ is the prediction error in the bth simulation iteration. We define Av.ResRMSPE similarly.

Table 2 summarizes the results which show that the VS-based predictions are much better than the standard analysis in almost all the cases. As we go from model (a) to model (c) the

TABLE 2
Prediction performance of the VS-based methodology and standard approach on varying-quality observations without any reference data

		VS		Std. app.	
Noise model	n	Av.RMSPE	Av.ResRMSPE	Av.RMSPE	Av.ResRMSPE
(a)	100	5.29 (4.04)	0.703 (0.22)	8.61 (15.37)	3.637 (1.82)
	500	4.046 (1.03)	0.281 (0.03)	4.826 (8.89)	4.416 (1.33)
	3000	3.927 (1.07)	0.141 (0.02)	3.228 (1.37)	5.306 (0.48)
(b)	100	9.67 (6.11)	1.796 (0.77)	37.38 (92.72)	14.717 (6.99)
	500	8.478 (5.04)	0.358 (0.07)	28.911 (75.28)	14.267 (7.56)
	3000	5.196 (3)	0.15 (0.02)	20.546 (33.01)	14.902 (8.07)
(c)	100	21.071 (11.29)	5.833 (1.39)	98.585 (206.89)	38.74 (19.03)
	500	26.325 (14.35)	1.376 (1.44)	66.6 (152.04)	36.354 (20.06)
	3000	13.722 (6.55)	0.23 (0.04)	94.429 (193.5)	31.606 (23.25)

prediction accuracy has compromised for both the VS-based as well as the standard approach with much higher impact for the later one. However, in terms of residual kriging efficiency the VS-based methodology is highly robust as compared to the ordinary kriging using the residuals obtained from o.l.s.

4.2. *With reference data.* In this subsection we consider a situation that is more similar to our case study. In addition to the n varying-quality observations in the hyperlocal region $\mathcal{R} = [0, \lambda_n]^2$, we have m-many high-quality observations available over a larger region $\mathcal{D} = [0, \Lambda_m]^2$. One example of the sampling points is shown in Figure 4(b). Our goal is to predict the process within the hyperlocal region $\mathcal{R}$ using the varying-quality observations. We again use a $4\lceil\lambda_n\rceil \times 4\lceil\lambda_n\rceil$ grid over the hyperlocal region of interest $\mathcal{R}$ to evaluate the predictions. In addition to the predictions obtained by the VS-based and standard methodology on the varying-quality observations, we also consider the global predictions obtained by using only the reference data on the larger region as shown in Figure 4(b). For this simulations we have considered the sample sizes for varying-quality observations to be equal to 50, 100 and 500 because the hyperlocal regions in our case studies do not contain very "large" (not more than 300) number of crowdsourced observations. For the reference data the sample sizes we have taken $m = 100$.

In Table 3, first, we compare the performances of the VS-based and standard predictions using hyperlocal noisy data based on RMSPE for both at the response level (Av.RMSPE) and residual level (Av.ResRMSPE). Clearly, we can see that VS-based predictions are uniformly better than the standard ones in all the considered cases. Next, we compare the VS-based predictions using hyperlocal noisy data and the predictions obtained by implementing the standard methodology on the high-quality reference data over a bigger region. We refer the later one as "Ref. Only." From Table 3 we see that, at response level (i.e., comparing Av.RMSPE) and under all noise models, the performance of the VS-based predictor using varying-quality observations is similar or slightly worse to the "Ref. Only" predictor when the number of hyperlocal noisy data and the high-quality reference data are comparable (i.e., the case when both n and $m = 100$.) In case we have larger sample size ($n = 500$) in the hyperlocal regions, we see a little gain in prediction efficiency in terms of Av.RMSPE. However, if we consider the residual kriging performance, that is, the ResRMSPE, the VS-based technique has outperformed the "Ref Only" kriging for all the cases, even when we have only $n = 50$ many varying-quality observations. As the kriging is more efficient when we have observations closer to the locations of our interest, the varying-quality hyperlocal observations along with the robust VS-based methodology improves the efficiency of the spatial prediction as compared to the corresponding "Ref. Only" version. Additional details regarding the simulation results, for example, the parameters of the models and choices of the regularity parameters, etc., are reported in Section C.1 of the Supplementary Material (Chakraborty, Lahiri and Wilson (2020)).

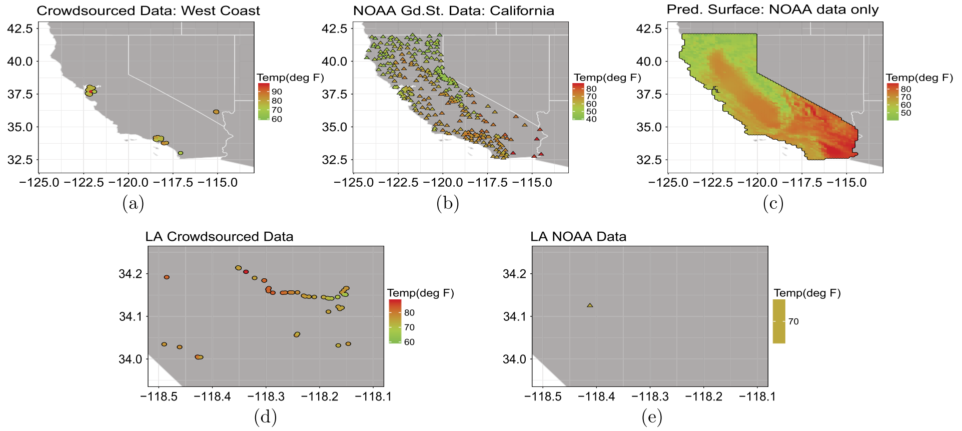
5. Case study: Spatial analysis of weather-signal data. In this section we analyze the WeatherSignal data described in Section 1.1 using the VS-based methodology (Section 3). Our goal for this noisy crowdsourced data set is to perform structure exploration and then prediction of the daily average ambient temperature process in hyperlocal regions of interest.

5.1. *Building hyperlocal prediction surfaces.* Here, we describe the VS-based analysis of the crowdsourced weather-signal data using the NOAA ground-station data as reference. We first select a hyperlocal region, as denoted by $\mathcal{R}$ in Section 2.3.1, around Los Angeles, CA, as shown in Figure 5(d). The analysis starts by defining a region large enough to have sufficient NOAA ground-station observations to build a reasonable global prediction surface

TABLE 3

Performance of hyperlocal predictions using the VS-based methodology, the standard approach and global predictions using reference data only. For these simulations we used reference data with sample size $m = 100$

		VS		Std. App.		Ref. Only	
Noise model	n	Av.RMSPE	Av.ResRMSPE	Av.RMSPE	Av.ResRMSPE	Av.RMSPE	Av.ResRMSPE
(a)	50	12.26 (12.71)	7.084 (5.08)	1740.696 (9518.03)	1745.244 (9604.81)	9.711 (8.54)	9.017 (7.46)
	100	10.877 (11.61)	6.104 (5.24)	230.117 (918.91)	224.56 (934.89)		
	500	8.787 (8.05)	6.287 (6.47)	358.694 (1976.29)	352.86 (1975.49)		
(b)	50	12.933 (13.1)	8.206 (6.76)	52,829.372 (662,485.91)	52,946.917 (664,727.7)		
	100	9.907 (10.81)	6.439 (5)	115.222 (923.6)	387.071 (915.98)		
	500	9.005 (8.66)	6.72 (5.06)	26.31 (19.18)	217.784 (15.88)		
(c)	50	12.33 (18.51)	8.7 (16.61)	10,198.908 (85,831.41)	9740.72 (85,082.65)		
	100	10.131 (10.93)	7.093 (5.04)	155.796 (126.08)	412.788 (31.68)		
	500	9.786 (8.45)	6.402 (5.24)	239.728 (29.49)	27.335 (8.35)		



(a) (b) (c)

(d) (e)

FIG. 5. (*a*) *Crowdsourced observations in CA*; (*b*) *Available ground-station observations*; (*c*) *Prediction surface using the standard approach on the ground-station data*; (*d*) *Crowdsourced observations in a hyperlocal region around Los Angeles*; (*e*) *ground-station observations in a hyperlocal region around Los Angeles*.

around the region of interest. In Figure 5(b) we plot the $m = 310$ ground-station observations in California. Using the standard approach on the NOAA ground-station data, as described in Section 3.1, we build a prediction surface for California and plot it in Figure 5(c). The model we use to estimate the mean is given by

$$\mu(\mathbf{s}) = \beta_0 + \beta_x s_x + \beta_y s_y + \beta_{xy} s_x s_y + \beta_h h(\mathbf{s}), \tag{5.1}$$

where $\mathbf{s} := (s_x, s_y)'$ and $h(\mathbf{s})$ denotes the elevation of the point $\mathbf{s}$. The mean model explains 79% (adjusted R^2) of the variability in the ground-station ambient temperatures in California.

We then fit a Matérn covariance to the observed residuals from the mean model estimation. Details of the variogram estimation are given in Table 4 and Figure 6. We then use standard kriging methodology with the estimated mean and covariance model to create the prediction surface $\{(\mathbf{s}, \hat{Y}(\mathbf{s})) : \mathbf{s} \in \mathcal{D}\}$, as shown in Figure 5(c).

As we can see in Figure 5(a), the spatial coverage of the crowdsourced data does not support a global prediction surface over California or even the coast of California. However, if we consider the 25 × 25 mile region ($\mathcal{R}$) in LA, as shown in Figure 5(d), the density of crowdsourced data is much higher as compared to only one ground-station observation (Figure 5(e)). While there is only one ground-station available at Los Angeles International Airport, the number of crowdsourced observations, $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$, in $\mathcal{R}$ is $n = 80$.

The next part of the analysis examines whether we can leverage the additional crowdsourced information through the VS-based methodology. We want to explore whether we can create a more reasonable and efficient prediction surface $\{(\mathbf{s}, \hat{Y}_{\mathrm{VS}}(\mathbf{s})) : \mathbf{s} \in \mathcal{R}\}$ over the region $\mathcal{R}$ in Los Angeles as compared to the surface obtained from the analysis of the ground-station data only, $\{(\mathbf{s}, \hat{Y}(\mathbf{s})) : \mathbf{s} \in \mathcal{R}\}$.

TABLE 4
Estimated Matérn parameters

Parameters	Estiamtes
partial sill (σ^2)	13.78
range (ρ)	0.36
nugget (τ^2)	7.95
smoothness (κ)	2.45

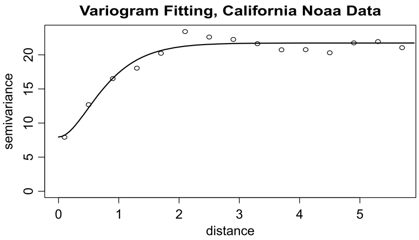


FIG. 6. *Variogram estimation.*

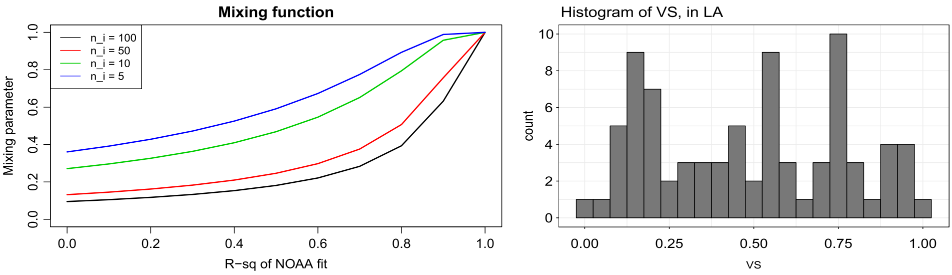
The VS-based analysis starts by computing the veracity score of the crowdsourced observations using the definition in equation (2.4). We set the baseline deviation $\alpha = 3$. In an ideal scenario, when the corresponding δ-neighborhood has very little variation and $\mathrm{IQR}(\boldsymbol{\xi}_i) \approx 0$, an observation with 3°F deviation from the corresponding benchmark value has a VS approximately equal to $\exp(-1) \approx 0.368$, while an observation with a 1°F deviation has a VS ≈ 0.716. To define the neighborhood for computation of the VS, we take $\delta = 0.08$ in the units of latitude and longitude. To choose a suitable mixing parameter ν, we use the function

$$\nu(\mathbf{s}_i) = 1 - \exp\biggl(\frac{-1}{(1 - \mathrm{R}^2)\sqrt{n(i)}}\biggr),$$

where R^2 is the adjusted R-squared for the estimation of the mean surface using NOAA ground-station data only and $n(i)$ is the number of crowdsourced data in the δ-neighborhood. As Figure 7(a) shows, this function is increasing in R^2 and decreasing in $n(i)$. $\nu(\mathbf{s}_i) = 1$ if $\mathrm{R}^2 = 1$ and $\nu(\mathbf{s}_i) = 0$ if $n(i) = \infty$. With this formulation the mixing parameter takes both the goodness of fit for the ground-station data and the number of crowdsourced observations used for local approximation of the target value into account. Using the specified parameters, we compute the VS for the crowdsourced observations in $\mathcal{R}$ and plot their empirical distribution in Figure 7(b).

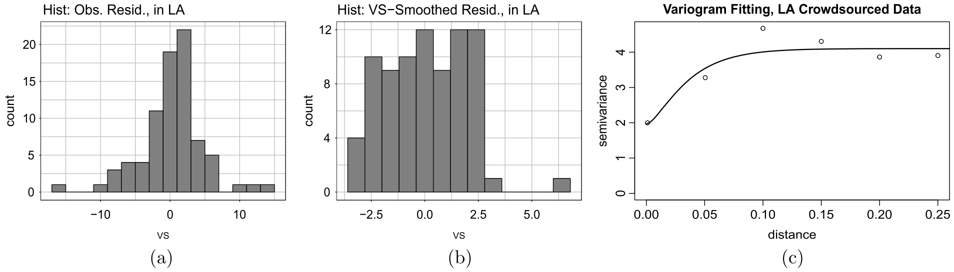
We next estimate the mean and covariance of the process. For robust estimation of the mean function, we use the weighted MM-type estimator, as discussed in Section 3.2, with the VS of the observations as the corresponding weights. Once the regression parameters are estimated, for a given smoothing parameter q in equation (3.4), we use the VS-based smoothing technique to reduce the effects of noise in the residual process as discussed in Section 3.3. Using the smoothed residuals, we estimate the covariance parameters and use the estimates to create a prediction surface using VS-based kriging as discussed in Section 3.4.

To make an optimal choice for q, we use the reference data. For a prespecified set of values of $q \in [0.05, 3]$, the covariance estimation and kriging are executed at the ground-station locations that are inside the hyperlocal region $\mathcal{R}$, and the q that minimizes the mean squared



(a) (b)

FIG. 7. *Mixing function* (*a*) *and the histogram of the veracity scores* (*b*) *for the crowdsourced observations in Los Angeles.*

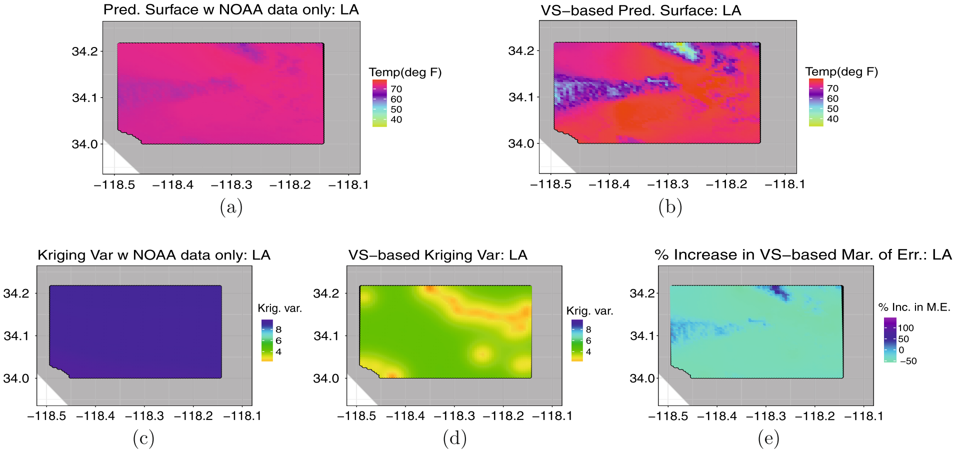


(a) (b) (c)

FIG. 8. *Histograms of the observed residuals* (*a*), *VS-based smoothed residuals* (*b*) *and the VS-based variogram fitting* (*c*) *for optimal* $q = 0.8$.

error of prediction at the stations is chosen to be optimal. In the analysis for the hyperlocal region around Los Angeles, there is only one station available, so we use the set of points with VS greater than or equal to 0.8 as test data and minimize leave-one-out cross-validated mean squared prediction error, that $n_*^{-1}\sum_j (Z(\mathbf{s}_j) - \hat{Y}_{\text{VS}}^{(-j)}(\mathbf{s}_j))^2$ where $\hat{Y}_{\text{VS}}^{(-j)}(\mathbf{s}_j)$ is the predicted value at $\mathbf{s}_j$ obtained using $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_{j-1}), Z(\mathbf{s}_{j+1}), \dots, Z(\mathbf{s}_n)\}$ as the training data and the sum is over the test data set whose cardinality is denoted by n_*. In Figures 8(a) and 8(b) we plot the histograms of the observed residuals from the VS-based robust regression and the residuals after the VS-based smoothing. The VS-based smoothing clearly reduces the spread of the residual values by smoothing out the large errors. In Figure 8(c) we show the robust variogram fitting of the VS-based smoothed residuals for the optimal choice of the smoothing parameter $q = 0.8$.

Given these analyses, we construct a prediction surface over the region $\mathcal{R}$ using equation (3.2). In Figure 9, we plot the hyperlocal prediction surfaces obtained by the standard analysis with the NOAA ground-station data only, as well as the one obtained by implementing the VS-based technique on the crowdsourced observations with the ground-station data as the reference. Clearly the prediction surface obtained from standard analysis of the ground-station data (Figure 9(a)) is too smooth to capture the local variability accurately. The prediction surface obtained by the VS-based analysis on crowdsourced data shows more variation across the space. To highlight the advantage of having crowdsourced observations, we compare the


Pred. Surface w NOAA data only: LA
Temp(deg F)
VS−based Pred. Surface: LA
Temp(deg F)
Kriging Var w NOAA data only: LA
Krig. var.
VS−based Kriging Var: LA
Krig. var.
% Increase in VS−based Mar. of Err.: LA
% Inc. in M.E.

(a) (b)

(c) (d) (e)

FIG. 9. (*a*) *Hyperlocal version of the same surface as in Figure* 5(*c*); (*b*) *Prediction surface obtained by the VS-based technique on the crowdsourced data in Los Angeles*; (*c*) *Residual kriging variance for the predictions using NOAA data only*; (*d*) *Residual kriging variance for the predictions using the VS-based predictions with crowdsourced data*; (*e*) *the % increase in the margin of error for the VS-based predictions as compared to the predictions with NOAA data.*

residual kriging variance surfaces in Figures 9(c) and 9(d). It is prominent from Figure 9(d) that the VS-based kriging variance is much smaller as compared to the global kriging using only the ground-station data, especially at locations that are close to the crowdsourced observations.

In addition, we illustrate the gain in efficiency by plotting the percentage increase in margin of error (at 95% confidence) for the VS-based predictions from the hyperlocal crowdsourced information as compared to the global prediction using ground-station data only, that is, $100 \times (\text{M.E.}(\hat{Y}_{\text{VS}}(\mathbf{s})) - \text{M.E.}(\hat{Y}(\mathbf{s})))/(\text{M.E.}(\hat{Y}(\mathbf{s})))$ where M.E. denotes the 'margin of error' (half of the length of the prediction interval) to predict the target response $Y(\mathbf{s})$. To compute the margin of error, we use ad hoc confidence intervals for the residual kriging predictor with ± 1.96 as the corresponding quantiles and then add the margin of error of the mean $(1.96 \times \text{s.e.}(\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}_{\text{VS}}))$ and the margin of error of the residual kriging predictor $(1.96 \times \sqrt{\text{Krig.Var.}(\tilde{\epsilon}(\mathbf{s}))})$. The margin of error for the standard predictor is computed similarly. A more theoretically justifiable interval can be obtained through spatial resampling technique as discussed in Lahiri (2003) but that requires further research and is beyond the scope of this study. In Figure 9(e) for most of the locations where the predictions have been carried out, there are decrease in the margin of errors for the VS-based predictions as compared to the global predictions using ground-station data only. At the locations that are close the crowdsourced observations, the VS-based prediction technique has achieved up to a 50% gain in efficiency.

The disadvantage of VS-based hyperlocal analysis is that the model is estimated very regionally and, hence, extrapolation of the estimated mean model outside the sample space is likely to give misleading and inefficient predictions. For example, in Figure 9(b) there are locations with elevations of more than 500 meters while the maximum elevation in the crowdsourced sample is 350 meters. This leads to poor predictions (e.g., ambient temperature less than 50°F) at some locations, as can be seen in Figure 9(e). Note that, though in those regions the efficiency of VS-based predictions falls short, the residual kriging variance (Figures 9(c) and 9(d)) for the VS-based kriging predictor is still less than the global kriging with NOAA data only. So, the loss in efficiency in VS-based predictions is solely due to the the extrapolation of the hyperlocally estimated mean function at points outside the covariate sample space.

We conduct a similar analysis for another hyperlocal region close to Brooklyn, NY and plot the results in Figure 10. The prediction surface in Figure 10(c) is obtained by using standard methodology on 120 ground-station observations over the East Coast, and the surface in Figure 10(d) is generated through VS-based hyperlocal analysis of the crowdsourced observations in Figure 10(b). Comparing these two prediction surfaces, we again see that the regional variation is prominent for the prediction surface obtained from VS-vased hyperlocal analysis whereas the global analysis generates a surface that is too smooth to accurately capture local variations. In Figure 10(f) the advantage of having crowdsourced data for hyperlocal prediction of the process is visible, as the kriging variance of the VS-based methodology is much smaller compared to Figure 10(e), especially in locations close to the crowdsourced observations. In Figure 10(g) we see up to 33% gain in margin of error by implementing the VS-based methodology on the crowdsourced data in locations close to the crowdsourced observations. Similar to the previous analysis of the Los Angeles data, the advantage of the VS-based hyperlocal predictions is lost if the predictions are attempted at locations too far from the crowdsourced observations or at locations with elevations outside the range of crowdsourced sample.

In addition to the VS-based hyperlocal analysis, we have also conducted the analysis for both of the hyperlocal regions in Los Angeles and Brooklyn with the standard approach without considering the veracity of the crowdsourced observations and then compared the

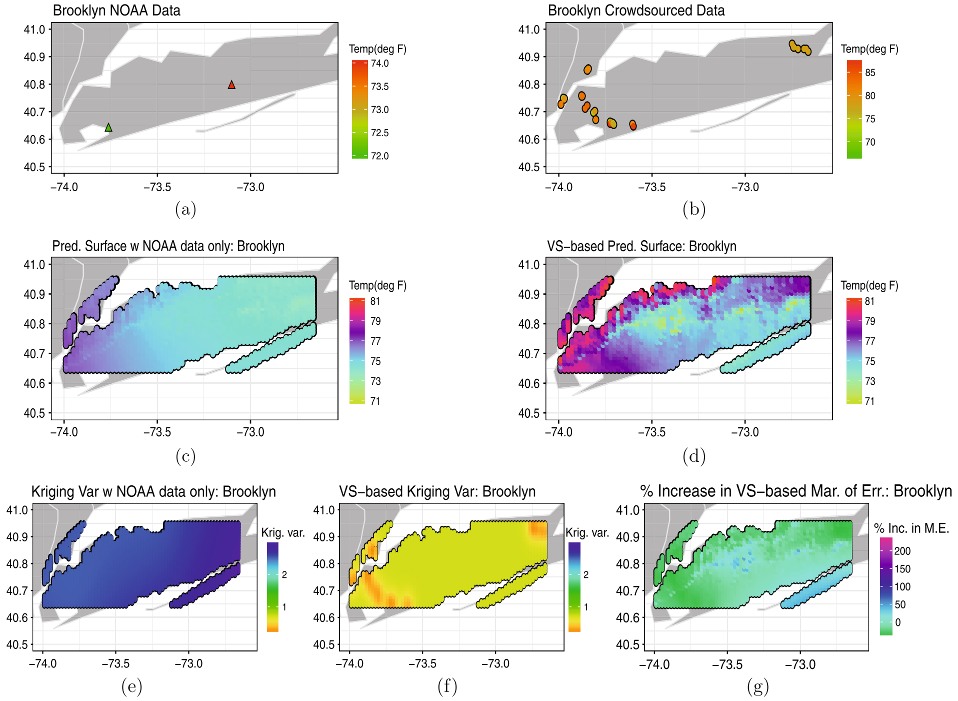

Brooklyn NOAA Data
Temp(deg F)
74.0
73.5
73.0
72.5
72.0
Brooklyn Crowdsourced Data
Temp(deg F)
85
80
75
70
41.0
40.9
40.8
40.7
40.6
40.5
−74.0
−73.5
−73.0
(a)
(b)
Pred. Surface w NOAA data only: Brooklyn
VS−based Pred. Surface: Brooklyn
Temp(deg F)
81
79
77
75
73
71
(c)
(d)
Kriging Var w NOAA data only: Brooklyn
VS−based Kriging Var: Brooklyn
% Increase in VS−based Mar. of Err.: Brooklyn
Krig. var.
2
1
% Inc. in M.E.
200
150
100
50
0
(e)
(f)
(g)

FIG. 10. (*a*) *Ground-station observations in the selected hyperlocal region*; (*b*) *Crowdsourced observations in the same region*; (*c*) *Prediction surface obtained by standard analysis of NOAA ground-station data*; (*d*) *Prediction surface obtained by the VS-based technique on the crowdsourced data*; (*e*) *Residual kriging variance for predictions using NOAA data only*; (*f*) *Residual kriging variances for the predictions using the crowdsourced data*; (*g*) *Percent increase in the margin of error for the VS-based predictions compared to the predictions with NOAA data*.

predictions with the global prediction surface obtained using reference data only. Comparing the plots in Figure 11 with Figure 9(e) and Figure 10(g) we can see that, in both Los Angeles and Brooklyn, the margins of error for the predictions using the standard approach are larger in all the locations as compared to the global predictions using ground-station data. In Brooklyn, even at the locations around the crowdsourced observations and with reference to the global prediction using ground-station data, the margin of error of standard predictions using the crowdsourced observations have increased by at the least 120%, whereas, as we have mentioned already, the VS-based methodology has achieved a decrease in the margin of error up to 33% (Figure 10(g)). Clearly, no gain from the “hyperlocal” analysis is achieved, as compared to the “global” prediction from the ground-station data, unless the robust VS-based methodology is employed on the varying-quality crowdsourced data.

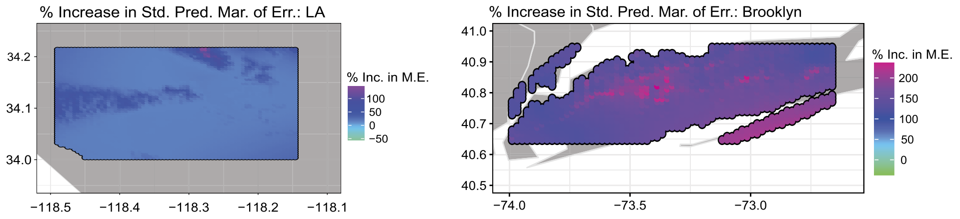


FIG. 11. *The increase in margin of error for the standard approach in hyperlocal regions in Los Angeles* (*left*) *and Brooklyn* (*right*).

5.2. *Validation at the ground stations.* The goal of the analysis in this section is to validate the predictions obtained by hyperlocal analysis of crowdsourced data using VS-based methodology. To do so, we have selected a set of 14 ground stations that satisfy the following criteria: (1) there are at least 30 crowdsourced data points available nearby with at least 20 observations with a VS greater or equal to 0.4; (2) the elevation of those stations is not too far from the range of the local crowdsourced samples. We have conducted 14 hyperlocal analyses, as described in Section 5.1, for hyperlocal structure exploration of the ambient temperature and then predicted at those selected ground-station locations to validate the VS-based predictions. We have omitted these 14 stations beforehand so that these are not used in defining the "benchmark" value at the crowdsourced data locations to compute VS; this way the validation data has no effect on the training phase of the predictions. We have also conducted the same hyperlocal analyses using the standard technique without taking the quality of the observations into account. The results are compiled in Table 5. The advantage of using the VS-based techniques as compared to the standard methodology is clear from the results. The RMSPE of the VS-based predictor for these 14 ground-stations is 3.71 while for the standard approach it is 4.54. More importantly, the average margin of error (at 95% confidence) for standard predictor is 13.61, and for the VS-based methodology it is 6.28. Relative to the standard methodology, on average, the VS-based technique has achieved approximately 54% gain in efficiency of the predictions.

6. Summary and conclusions.

In this paper we have introduced the veracity score to assess the quality of observations in geostatistical settings. The VS is defined by comparing the varying quality observations with a benchmark. We used the ground-station data as our reference to define the benchmark values in the case studies. The similar scoring approach to assess the veracity of the observations can be used in other contexts as well. We have also discussed the case when no other reference information is available and propose a version of VS using locally and robustly estimated measure of center as the benchmark. A robust approach for modeling varying-quality spatial data using the VS has been proposed and evaluated. We have illustrated the VS-based methodology on a crowdsourced data set coming from the mobile app WeatherSignal using NOAA ground-station data as the reference. Both the simulation studies in Section 4 and the case studies in Section 5.1 show the advantages of the VS-based methodology over the standard geostatistical approach when dealing with noisy spatial data. In addition, by implementing the VS-based methodology on the varying-quality local crowdsourced data we can achieve a more accurate and efficient hyperlocal predictions as compared to the global prediction obtained from the analysis of ground-station data only.

In the analysis of crowdsourced data using the VS-based methodology, the model is estimated using observations in a hyperlocal region. Predicting at more distant locations or with covariates outside the range of the sample may provide misleading predictions, as we have seen for some of the locations in Figure 9(b) and Figure 10(d). The mean and covariance models used to explore the structures of the average temperature process are quite simple, yet reasonable and effective for hyperlocal analysis of ambient temperature. More complex models like nonlinear regression models (Frei (2014)) and anisotropic covariance (Haskard (2007)) can be incorporated in the VS-based technique to increase flexibility of the analysis. The VS-based kriging automatically reduces the impact of the corrupted observations and thus, it does not require removing the outliers manually (e.g., see Frei (2014)) which is often not feasible when dealing with large crowdsourced spatial data. In addition, as the veracity of the observations has been measured nonparametrically using "local" summaries, the proposed VS-based kriging does not require any distributional assumption (e.g., Gaussian, see Lussana, Uboldi and Salvati (2010)) on the underlying spatial process or the noise associated with it. The analysis presented in this paper shows that the systematic incorporation of

TABLE 5

Predictions using both the VS-based and standard approach at the ground stations with crowdsourced observations in proximity

STATION_NAME	Target temp.	PredTemp.VS	VS.ME	PredTemp.Std	Std.ME
CHICAGO OHARE INTERNATIONAL AIRPORT IL US	76	76.01	6.22	76.75	8.74
WASHINGTON DULLES INTERNATIONAL AIRPORT VA US	79	82.80	5.13	75.33	18.35
WASHINGTON REAGAN NATIONAL AIRPORT VA US	80	81.95	7.77	75.52	31.64
MIAMI INTERNATIONAL AIRPORT FL US	79	77.81	0.50	78.57	1.23
LITTLE TUJUNGA CALIFORNIA CA US	68	64.78	6.01	63.74	7.41
LOS ANGELES INTERNATIONAL AIRPORT CA US	68	68.91	3.31	67.87	4.26
BEVERLY HILLS CALIFORNIA CA US	70	67.94	6.27	68.12	7.54
TOLEDO EXPRESS AIRPORT OH US	75	79.45	5.72	79.39	8.18
DETROIT METROPOLITAN AIRPORT MI US	76	78.79	6.66	80.74	9.73
MINNEAPOLIS ST PAUL INTERNATIONAL AIRPORT MN US	70	77.03	6.67	76.97	10.96
CARLOS AVERY MINNESOTA MN US	69	73.33	11.46	74.65	19.05
JFK INTERNATIONAL AIRPORT NY US	72	78.24	3.06	80.82	3.82
ISLIP LI MACARTHUR AIRPORT NY US	74	75.10	6.29	75.61	8.79
AUSTIN BERGSTROM INTERNATIONAL AIRPORT TX US	81	78.87	2.24	86.50	10.19

VS in the geostatistical analysis helps us capture the local variability of the ambient temperature field by considering crowdsourced data in hyperlocal regions. The VS-based kriging decreases the margin of prediction errors up to 50% as compared to the global predictions from ground-station data only. On the other hand, if the same analysis is carried out on the noisy crowdsourced data with standard kriging, there is no gain in efficiency. In fact, there are locations, even close to the crowdsourced observations, where the margin of prediction errors by standard methods are more than 80% higher than the corresponding global predictions.

There are several interesting future directions for this work. First, we have not provided theoretical justification for the superiority of the VS-based methodology as compared to the standard approach in the analysis of noisy spatial data. Inspired by the simulations executed in this work, we believe that under a suitable spatial asymptotic framework (e.g., *mixed-increasing domain*, Hall and Patil (1994); Lahiri, Lee and Cressie (2002)) and a fairly general nonstationary noise model (e.g., the *additive-multiplicative* model defined in equation (2.3)), we can theoretically justify the robustness and efficiency of the VS-based methodology (for details, see Chakraborty and Lahiri (2019)). Second, the methodology discussed in this article can be systemically extended to develop a more sophisticated VS-based kriging technique that incorporates both the ground-station data and the crowdsourced data for spatial prediction. Third, a spatiotemporal VS and corresponding methods for real-time crowdsourced data can be developed by considering neighborhoods in both space and time.

Acknowledgments. This material is based upon work supported in whole or in part with funding from the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government. Special thanks are extended to the OpenSignal team that was in charge of the academic partnership program in 2015 for making the data available to the authors. The authors also thank the Editor, the Associated Editor and three anonymous referees for a number of thoughtful comments that significantly improved the paper.

This work was supported in part by NSF Grant DMS-1613192.

SUPPLEMENTARY MATERIAL

Supplement to "A statistical analysis of noisy crowdsourced weather data" (DOI: 10.1214/19-AOAS1290SUPP; .pdf). This file contains additional details on data preprocessing, the simulations, and the case study. It contains additional plots, tables and discussions to support our claims and findings in the main article.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. Dover, New York.

ACCUWEATHER (2015). AccuWeather launches AccUcast, providing exclusive crowdsourced weather feature worldwide. Available at https://www.accuweather.com/en/press/50601069. Accessed: 2019-01-30.

ALLAHBAKHSH, M., BENATALLAH, B., IGNJATOVIC, A., MOTAHARI-NEZHAD, H. R., BERTINO, E. and DUSTDAR, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Comput*. **17** 76–81.

CHAKRABORTY, A. and LAHIRI, S. N. (2019). On statistical properties of a veracity scoring method for spatial data. arXiv preprint arXiv:1906.08843.

CHAKRABORTY, A., LAHIRI, S. N. and WILSON, A. (2020). Supplement to "A statistical analysis of noisy crowdsourced weather data." https://doi.org/10.1214/19-AOAS1290SUPP.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.

CRESSIE, N. and HAWKINS, D. M. (1980). Robust estimation of the variogram. I. *J. Int. Assoc. Math. Geol.* **12** 115–125. MR0595404 https://doi.org/10.1007/BF01035243

DALTON, A. (2016). Dark Sky's hyperlocal weather app is now available on the web. Available at https://www.engadget.com/2016/09/20/dark-sky-hyperlocal-weather-app-desktop-web/. Accessed: 2019-01-30.

FLORIO, E. N., LELE, S. R., CHANG, Y. C., STERNER, R. and GLASS, G. E. (2004). Integrating AVHRR satellite data and NOAA ground observations to predict surface air temperature: A statistical approach. *Int. J. Remote Sens*. **25** 2979–2994.

FREI, C. (2014). Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol*. **34** 1585–1605.

GANDIN, L. S. (1988). Complex quality control of meteorological observations. *Mon. Weather Rev*. **116** 1137–1156.

GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.

GENTON, M. G. (1998). Highly robust variogram estimation. *Math. Geol*. **30** 213–221. MR1610687 https://doi.org/10.1023/A:1021728614555

GHOSH, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Stat*. **42** 1957–1961. MR0297071 https://doi.org/10.1214/aoms/1177693063

GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554 https://doi.org/10.3150/12-BEJSP06

HALL, P. and PATIL, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probab. Theory Related Fields* **99** 399–424. MR1283119 https://doi.org/10.1007/BF01199899

HARRIS, P., BRUNSDON, C., CHARLTON, M., JUGGINS, S. and CLARKE, A. (2014). Multivariate spatial outlier detection using robust geographically weighted methods. *Math. Geosci*. **46** 1–31. MR3158063 https://doi.org/10.1007/s11004-013-9491-0

HASKARD, K. A. (2007). An anisotropic Matérn spatial covariance model: REML estimation and properties, Ph.D. thesis, Univ. Adelaide.

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

KOLLER, M. and STAHEL, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Comput. Statist. Data Anal*. **55** 2504–2515. MR2787008 https://doi.org/10.1016/j.csda.2011.02.014

KÜNSCH, H. R., PAPRITZ, A., SCHWIERZ, C. and STAHEL, A. W. (2011). Robust estimation of the external drift and the variogram of spatial data. In *ISI* 58*th World Statistics Congress of the International Statistical Institute*, *Dublin*, *Ireland* 21–26.

LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York. MR2001447 https://doi.org/10.1007/978-1-4757-3803-2

LAHIRI, S. N., LEE, Y. and CRESSIE, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *J. Statist. Plann. Inference* **103** 65–85. MR1896984 https://doi.org/10.1016/S0378-3758(01)00198-7

LARK, R. M. (2000). A comparison of some robust estimators of the variogram for use in soil survey. *Eur. J. Soil Sci*. **51** 137–157.

LORENC, A. C. (1986). Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc*. **112** 1177–1194.

LUSSANA, C., UBOLDI, F. and SALVATI, M. R. (2010). A spatial consistency test for surface observations from mesoscale meteorological networks. *Q. J. R. Meteorol. Soc*. **136** 1075–1088.

MATHERON, G. (1962). *Traité de géostatistique appliquée*, *Tome I. Memoires du BRGM* (*Paris*) **14**. Technip, Paris.

MOYNIHAN, T. (2015). Clever app turns everyone into a roving weather reporter. Available at https://www.wired.com/2015/10/clever-app-turns-everyone-roving-weather-reporter/. Accessed: 2019-01-30.

PAPRITZ, A. (2018a). Tutorial and manual for geostatistical analyses with the R package georob. Available at https://cran.r-project.org/web/packages/georob/vignettes/georob_vignette.pdf. Accessed: 2019-02-12.

PAPRITZ, A. (2018b). georob: Robust geostatistical analysis of spatial data. R package version 0.3-7.

SEN, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Stat*. **39** 1724–1730. MR0232522 https://doi.org/10.1214/aoms/1177698155

SOSKO, S. and DALYOT, S. (2017). Crowdsourcing user-generated mobile sensor weather data for densifying static geosensor networks. *ISPRS Int.l J. Geo-Inf*. **6** 61.

SUN, S. and LAHIRI, S. N. (2006). Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā* **68** 130–166. MR2301568

THORNTON, P. E., RUNNING, S. W. and WHITE, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol*. **1905** 214–251.

TODOROV, V. and FILZMOSER, P. (2009). An object-oriented framework for robust multivariate analysis. *J. Stat. Softw*. **32** 1–47.

VANCUTSEM, C., CECCATO, P., DINKU, T. and CONNOR, S. J. (2010). Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa. *Remote Sens*. *Environ*. **114** 449–465.

WILLET, J. B. and SINGER, J. D. (1988). Another cautionary note about R^2: Its use in weighted least-square regression analysis. *Amer*. *Statist*. **42** 236–238.