Spatial Sampling Strategies with Multiple Scientific Frames of Reference

Paul B. Reverdy
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104
preverdy@seas.upenn.edu

Thomas F. Shipley
Department of Psychology
Temple University
Philadelphia, PA 19122
tshipley@temple.edu

Daniel E. Koditschek
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104
kod@seas.upenn.edu

Abstract

We study the spatial sampling strategies employed by field scientists studying aeolian processes, which are geophysical interactions between wind and terrain. As in geophysical field science in general, observations of aeolian processes are made and data gathered by carrying instruments to various locations and then deciding when and where to record a measurement. We focus on this decision-making process. Because sampling is physically laborious and time consuming, scientists often develop sampling plans in advance of deployment, i.e., employ an offline decision-making process. However, because of the unpredictable nature of field conditions, sampling strategies generally have to be updated online. By studying data from a large field deployment, we show that the offline strategies often consist of sampling along linear segments of physical space, called *transects*. We proceed by studying the sampling pattern on individual transects. For a given transect, we formulate model-based hypotheses that the scientists may be testing and derive sampling strategies that result in optimal hypothesis tests. Different underlying models lead to qualitatively different optimal sampling behavior. There is a clear mismatch between our first optimal sampling strategy and observed behavior, leading us to conjecture about other, more sophisticated hypothesis tests that may be driving expert decision-making behavior.

Keywords: Spatial sampling, frames of reference, representation, scientific decision making

Acknowledgements

This work was supported in part by NSF INSPIRE Award # 1514882.

1 Introduction

Aeolian processes, which couple the geophysical interactions between wind and terrain, are driving forces behind the production of atmospheric dust, erosion of soils, and the evolution of sand dunes in deserts [6]. These phenomena are particularly important in agriculture and climate science. Many environmental factors are relevant (wind turbulence, topography, vegetation, etc.), which leads to complex physics that are difficult to reproduce in the laboratory. Therefore, studying aeolian processes necessitates field work where scientists make observations and take measurements. There is growing reason to expect that advances in robotics [4] will soon yield a class of small legged machines capable of carrying the instruments required for these purposes of aeolian science [8]. This has prompted us to explore the prospects for developing formal models of the aeolian data collection enterprise such as might better enable human scientists to usefully direct the activities of such robotic field assistants.

A deep understanding of aeolian processes in the field remains at a relatively early stage of development; scientists have imprecise priors about what to expect when planning a field campaign. In this situation the ideal measurements would cover an area of interest at a density that would allow observation of relevant spatial variation at any scale. In the absence of infinite resources such fine-scale uniform spatial coverage of their area of interest is not practical [1]. Limited field time and financial resources will determine coverage, so a key decision-making process at the heart of field work is deciding how to allocate a limited sampling budget.

In this paper we consider sampling location data from an ongoing large-scale field campaign in Oceano, California [7]. The Oceano Dunes state park, located on the coast between Santa Barbara and San Luis Obispo, is open to the public for riding all-terrain vehicles (ATVs), which produce atmospheric dust. The field campaign was sponsored by the local air quality board in response to complaints about how ATV-generated dust was affecting local air quality. Scientists from the Desert Research Institute (DRI) were commissioned to study the processes by which dust was being generated and possible policies designed to mitigate the dust production [7, Appendix B].

The area of interest covered approximately 15 km^2, and the initial August 2013 field campaign gathered 360 samples (969 samples total over the six field campaigns in the data set). If the scientists had aimed for uniform coverage on a grid in August 2013 they would have resulted in an inter-sample distance of approximately 200 meters. However, this level of resolution is insufficient for identifying the role of the geomorphology, as it permits less than one sample per dune, whose spatial extent is on the order of 200-300 meters. Furthermore, difficulties in traversing the terrain mean that the identical amount of field resources dedicated to a uniform sampling strategy would likely have achieved fewer than 360 samples with a corresponding further decrease in resolution. Therefore, the decision makers (DMs) decided to focus on two types of individual transects, which they could sample at a higher spatial frequency. One type was oriented parallel to the prevailing wind direction (120°), and the other type was oriented parallel to the shore and thus also to the dune structures (0°). These two frames were not orthogonal due to the non-orthogonality of the two relevant orientations defined by the two physical drivers: wind direction and shore topography.

2 Field data

Figure 1 shows sampling locations from the six field campaigns conducted as part of the DRI Oceano Dunes study between 2013 and 2016. The data clearly show many linear patterns of sampling locations, which the field scientists refer to as linear *transects*. Sampling resources were focused in transects which raises three questions: 1) why were these orientations selected; 2) why were the transects located where they were; and 3) why was a regular sampling pattern used on the transects? We conducted interviews with Vicken Etyemezian, one of the senior scientists in charge of the study, to understand the decision-making process associated with these three questions. The results of the interviews are as follows.

2.1 Orientation

The initial sampling study in August 2013 was designed to measure baseline information about the rates of dust production in different areas in the park. The scientists hypothesized that there would be spatial gradients in the rates of dust production as measured by the PI-SWERL device [2]. As aeolian processes are driven by the wind, they further hypothesized that the spatial gradients would be strongly oriented along the prevailing wind direction. Therefore, the East-West transects are oriented along the prevailing wind direction and sampled at an approximately uniform 100 meter spatial resolution, selected as a compromise between achieving high spatial resolution and covering large amounts of terrain. The North-South spacing (approximately 300 meters) of these transects reflects an attempt to gather coarse North-South gradient information, while the transects oriented North-South are designed to more formally capture gradient information along a linearly independent direction [1].

2.2 Location

The choice of location for transects is influenced by many factors, including the availability of prior meteorological measurements, land access rights, existence of vegetation, and the difficulty of traversing the physical terrain. The long 2013 transect just south of the center of the figure was designed to provide a baseline measurement of dust production upwind of the reference point for the air quality district, while the other transects were designed to provide control information in areas that were off-limits to ATV riding.

2.3 Within-transect sampling

Sampling along a given transect was done approximately every 100 meters by pre-programming GPS coordinates into a handheld receiver and then attempting to take measurements at the closest-possible location. Sometimes, as can be seen in the sampling locations at the northern edge of the park, the physical terrain made it difficult to take measurements in the desired locations.

The spatial frequency of measurement was closely related to the spatial frequency (approximately 200-300 meters) of the dune field; the resolution was chosen to ensure that there would be several measurements on each dune. In this way, the scientists hoped to capture spatial gradients that were associated with the dune topography. During a given field campaign, the sampling strategy was generally held fixed (i.e., the sampling locations were determined before going into the field). Significant changes to the sampling strategy were only made between field campaigns.

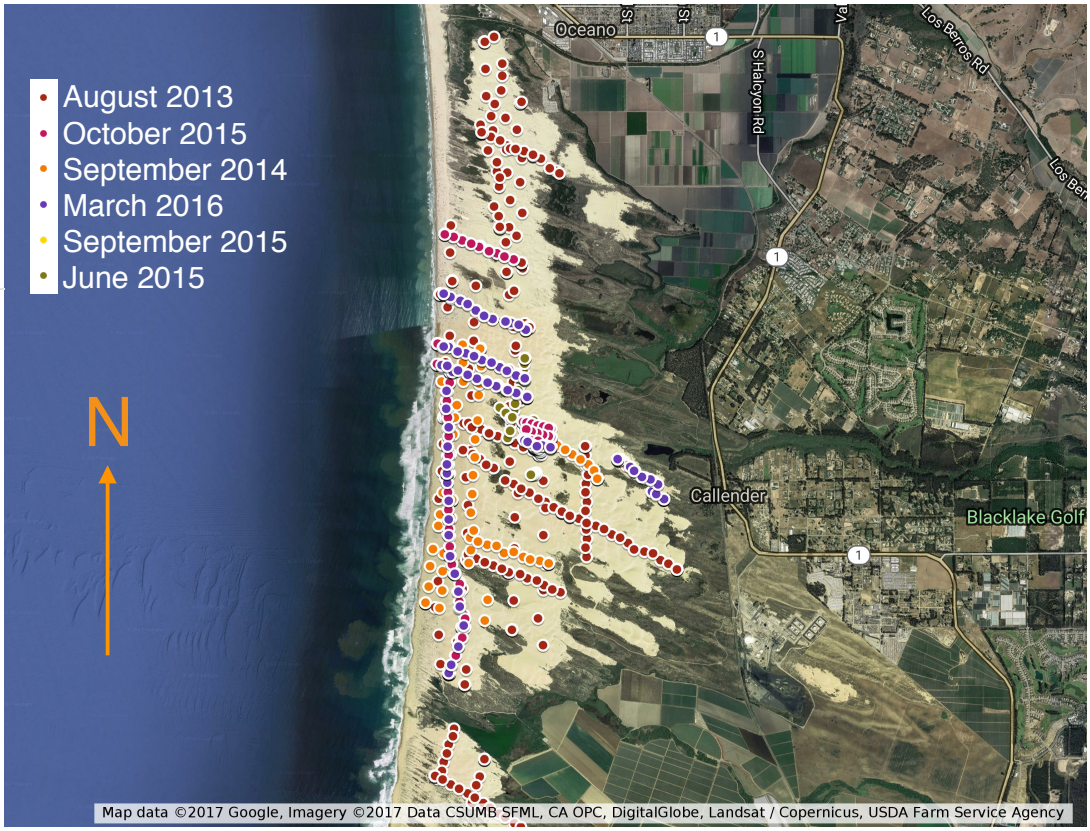

Figure 1: Sampling locations from the six field campaigns associated with the DRI Oceano study, broken out by field campaign date. Note the clear linear patterns of sampling locations that define transects. Many of the transects have strong East-West orientations, designed to match the heading of the prevailing winds. Other transects are strongly North-South, designed to capture gradient information along a direction "orthogonal" to the prevailing wind direction. Data courtesy of Vicken Etyemezian, Desert Research Institute.

3 Hypothesis testing on a transect

In the previous section we showed that, in this study, the field scientists focused their sampling on linear transects aligned with two reference frames defined by relevant physical processes. In this section, we consider the problem of deciding

where to sample on a given transect. We postulate that the decision-making process is guided by the desire to perform statistical hypothesis testing on the resulting data and investigate the spatial sampling patterns induced by maximizing the statistical power of tests associated with two different underlying models of the observed data. By comparing the induced to the observed sampling patterns, we can begin to probe the types of hypotheses the scientists are implicitly making about the structure of their data.

We restrict ourselves to a given linear transect of length ℓ and consider the problem of deciding where to sample on that interval. Let $x \in [0, \ell]$ represent locations on the transect. We suppose that the scientist is taking measurements of some quantity whose true value at location x is given by an unknown function $f(x)$. Furthermore, we assume that the measurement device produces observations

$$y = f(x) + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

which are the true function value corrupted by zero-mean Gaussian noise with known variance σ^2.

In interviews, Dr. Etyemezian emphasized that a major goal of the sampling project was to determine the extent to which quantities thought to contribute to dust production varied with different environmental covariates, such as distance from the ocean (i.e., along a wind-oriented transect) and policy choices, such as whether or not ATV riding was permitted in a given area. This qualitative question can be cast quantitatively in terms of several different types of statistical hypothesis tests, including change point detection and tests on the coefficients of linear and nonlinear regressions. The data produced by the PI-SWERL device is quite noisy, likely because of high local variability in the quantity it is designed to measure [1]. Therefore, we regard it as unrealistic that a field campaign would be designed to perform a change point detection, as any true change point would be difficult to detect in the presence of measurement noise.

Having rejected a simple change point detection test, we consider hypothesis testing on the parameters of a regression model. In the interest of parsimony, we restrict ourselves to linear regression models. As a first hypothesis, suppose that the decision maker assumes that the unknown function $f(x)$ is linear in x:

$$f(x) = mx + b. \tag{2}$$

When there are N measurements y_i made at locations x_i for $i \in \{1, \ldots, N\}$ fitting the function (2) is an ordinary least squares (OLS) problem

$$y = X\beta$$

where $\beta = [m, b]^T$ and $X = [x_1, \ldots, x_N; 1, \ldots, 1]^T$ is the $N \times 2$ matrix of regressors. The qualitative question of whether or not f varies with x then reduces to the quantitative question of whether or not the slope m is equal to zero.

3.1 Slope test

For the linear model (2), a natural question to ask is the following: is the slope m statistically different from zero? Such a question can be answered by conducting a hypothesis test on the OLS coefficients. We adopt a frequentist statistical framework and design a sampling strategy to maximize the statistical power of the frequentist hypothesis test on m, the slope coefficient.

Assuming that the Gaussian noise ε is independent across observations i, the Gauss-Markov theorem [3] holds and the estimated regression coefficients $\hat{\beta}$ are themselves Gaussian distributed with mean β and variance $\sigma^2(X^TX)^{-1}$:

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X^TX)^{-1}\right).$$

Maximizing the power of the hypothesis test is equivalent to minimizing the variance of the estimator $\hat{\beta}$. The matrix $(X^TX)^{-1}$ can be expressed as

$$(X^TX)^{-1} = \frac{1}{n}\frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}, \tag{3}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$. A bit of additional algebra then shows that the variance of $\hat{\beta}_2$, the estimator of the slope coefficient m, is given by $\sigma^2_{\hat{\beta}_2} = 1/\sum_i (x_i - \bar{x})^2$, which is the inverse of the moment of inertia of the points x_i.

The solution to the problem of maximizing the moment of inertia $\sum_i (x_i - \bar{x})^2$ on the interval $[0, \ell]$ is simple: pick $N/2$ points at either end of the interval. Clearly, the uniform spatial sampling observed in the field is inconsistent with this solution, so we reject the hypothesis that the scientists are performing a simple slope test and seek a more sophisticated sampling criterion.

3.2 Fourier analysis

Having rejected the simple slope test we consider a more complex hypothesis test arising from linear regression. In conversation, the scientists implied that they first wanted to eliminate the possibility that there was more complex spatial

behavior before testing the hypothesis that a linear model fit had zero slope, as the dune topography suggests that there may be some spatially-periodic variation in the measured variables. OLS tools can again be helpful here in the form of Fourier analysis. This is equivalent to assuming the following model for $f(x)$:

$$f(x) = \sum_{k=1}^{M} a_k \cos\left(\frac{2\pi k x}{N}\right) + \sum_{k=1}^{M} b_k \sin\left(\frac{2\pi k x}{N}\right). \tag{4}$$

Suppose that we have N uniformly-spaced samples at $x_i = i\lambda, i \in \{0, 1, \ldots, N-1\}$, where $\lambda = \ell/(N-1)$ is the spatial frequency. The well-known Nyquist sampling condition implies that we must have $M < N/2$, which provides an upper bound on the spatial frequencies that can be captured with a given sampling strategy: effectively, with a sampling frequency of λ, the highest spatial frequency that can be captured is $\lambda/2$.

Denote by $y_i, i \in \{0, 1, \ldots, N-1\}$ the N measurements of $f(x)$ taken at $x_i = i\lambda, i \in \{0, 1, \ldots, N-1\}$. Then the OLS estimators for the amplitudes a_k and b_k are the discrete Fourier transform coefficients

$$\hat{a}_k = \frac{2}{N} \sum_{i=0}^{N-1} y_i \cos\left(\frac{2\pi k i}{N}\right), \quad \hat{b}_k = \frac{2}{N} \sum_{i=0}^{N-1} y_i \sin\left(\frac{2\pi k i}{N}\right)$$

and the covariance matrix is $C = \frac{2\sigma^2}{N} I_{2M}$, which permits hypothesis testing [5, Section 4.4].

The primary experimental design decision to be made in this context is the selection of the sampling spatial frequency λ, which determines the highest spatial frequency detectable in the model. Thus, the choice of sub-dune-length-scale spatial sampling is consistent with a desire to measure spatial processes on the order of the dune length scale. Further careful investigation would be required to verify the extent to which the spatial sampling frequency is greater than twice the dune length scale, as required by the Nyquist criterion.

4 Conclusion

We investigated the spatial sampling behavior exhibited by scientists conducting research in the field. Records of field campaigns and personal interviews show that sampling tended to be performed along linear segments called *transects* and that these transects were oriented to coincide with physical drivers of the processes under study, namely the prevailing wind direction and the shoreline. We postulated that the pattern of sampling within a given transect could reveal implicit assumptions that the scientists made about the spatial structure of the scientific phenomena.

We formulated statistical hypothesis tests that might be driving the sampling behavior and considered the spatial pattern of sampling that would maximize the statistical power of the associated hypothesis test. There was a clear qualitative difference between the observed sampling behavior and that associated with a simple slope test, so we rejected that model and considered a Fourier series model. Maximizing the statistical power of a test associated with the coefficients of the Fourier series results in a regular pattern of sampling that is qualitatively similar to the observed pattern. Ongoing work will further investigate the degree to which the sampling behavior under the Fourier series model matches the observed behavior. We will conduct interviews to discover if this quantitative representation is consistent with the scientists' implicit assumptions about spatial structure and introduce other candidate models as may be suggested by the comparisons.

References

[1] V. Etyemezian. Personal communication, February 2017.

[2] V. Etyemezian, G. Nikolich, S. Ahonen, M. Pitchford, M. Sweeney, R. Purcell, J. Gillies, and H. Kuhns. The portable in situ wind erosion laboratory (PI-SWERL): A new method to measure PM 10 windblown dust properties and potential for emissions. *Atmospheric Environment*, 41(18):3789–3796, 2007.

[3] A. S. Goldberger. *A Course in Econometrics*. Harvard Univ. Press, 1991.

[4] G. C. Haynes, J. Pusey, R. Knopf, A. M. Johnson, and D. E. Koditschek. *Laboratory on legs: an architecture for adjustable morphology with legged robots*, volume 8387, pages 83870W–14. SPIE, 2012.

[5] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I : Estimation Theory*. Prentice Hall, 1993.

[6] N. Lancaster, A. Baas, and D. Sherman. Volume 11, Aeolian geomorphology. In J. F. Shroder, editor, *Treatise on Geomorphology*. Academic Press, San Diego, 2013.

[7] S. of California. Oceano dunes state vehicular recreation area, rule 1001 draft particulate matter reduction plan. Technical report, Department of Parks and Recreation, Oceano State Park, 2013.

[8] F. Qian, D. Jerolmack, N. Lancaster, G. Nikolich, P. Reverdy, F. Roberts, Sonia, T. Shipley, R. S. Van Pelt, T. M. Zobeck, and D. Koditschek. Ground robotic measurement of aeolian processes. *Aeolian Research*, page (under review), 2016.