On Geodetic Distance Computations in Spatial Modeling

Sudipto Banerjee

Division of Biostatistics, School of Public Health, University of Minneapolis,
Minnesota 55455, U.S.A.
email: sudiptob@biostat.umn.edu

Summary. Statisticians analyzing spatial data often need to detect and model associations based upon distances on the Earth's surface. Accurate computation of distances are sought for exploratory and interpretation purposes, as well as for developing numerically stable estimation algorithms. When the data come from locations on the spherical Earth, application of Euclidean or planar metrics for computing distances is not straightforward. Yet, planar metrics are desirable because of their easier interpretability, easy availability in software packages, and well-established theoretical properties. While distance computations are indispensable in spatial modeling, their importance and impact upon statistical estimation and prediction have gone largely unaddressed. This article explores the different options in using planar metrics and investigates their impact upon spatial modeling.

Key words: Correlation functions; Geodetic distances; Geographical Information Systems; Isotropic models; Map projections; Spatial range; Spherical coordinates.

1. Introduction

The analysis and modeling of geographically referenced data play an indispensable role in diverse disciplines such as environmental sciences, ecology, and public health. Such data are often obtained from a set of locations referenced by geographical coordinates (longitude and latitude) that form the "spatial domain." Spatial modeling attempts to detect and model associations between the observed variables as a function of distances (and perhaps angles) between locations.

Distance computations are indispensable in spatial analysis. Precise intersite distance computations are used in variogram analysis to assess the strength of spatial association. They help in specifying priors on the range parameter in Bayesian modeling (Ecker and Gelfand, 1997), and in setting starting values for the nonlinear least-squares algorithms in classical analysis (Cressie, 1993), making them crucial for correct interpretation of spatial range and the convergence of statistical algorithms. Yet, this is not an issue that has received much attention in the existing statistical literature and ambiguity prevails among practicing statisticians about distance metrics. For example, the analysis of the scallops data appearing in Kaluzny et al. (1998, p. 76–79), and in Ecker and Gelfand (1997), uses naive Euclidean distances treating the geographical coordinates as planar. Except when the spatial domain is "small enough" as to have negligible curvature (where how "small" is "small enough" depends upon the specific application), the usual planar metrics for calculating distances are inappropriate. Treating geodetic coordinates as planar can induce deceptive anisotropy in the models because of the difference in differentials in longitude and latitude (a unit increment in degree longitude is not the same length as a unit increment in degree latitude except at the equator).

Spurious nonstationarity may be induced as well due to the systematic properties of these differentials.

Nevertheless, Euclidean metrics are popular due to their simplicity and availability in standard software. More importantly, statistical modeling of spatial correlations proceed from *correlation functions* that are often valid only with Euclidean metrics. As we demonstrate later, applying these metrics on geographical coordinates requires care, and can otherwise have unattractive consequences on statistical estimation and subsequent interpretation. Note that in geostatistics interest often resides in points that are "closer" together, so the sensitivity of planar metrics may seem irrelevant. However, an important feature of formal spatial modeling (particularly isotropic models) is inference on the *effective spatial range*, a critical *distance* beyond which spatial correlation is deemed negligible. The range is relative to the spatial domain and is likely more sensitive to the definition of distance, especially for larger domains.

This article explores options for computing distances and investigates their impact upon statistical modeling, keeping in mind the practicing modeler. While mathematical cartography presents a rich literature (see, e.g., Snyder, 1987) in the study of geodetic distortions and planar projections, such topics focus upon the *thematic* properties based upon mapping objectives, but are less useful for practicing statisticians seeking appropriate intersite distances for statistical modeling.

Spatial statisticians, however, often need to use cartographic concepts and do so using Geographical Information Systems (GIS). These databases offer versatile interfaces for manipulating and visualizing spatial data and play an indispensable role in spatial statistics that is too huge to

be addressed comprehensively here (see, e.g., Jones, 1997). Focusing upon distance computations, GIS offers a wide array of planar map projections using appropriate coordinate transformations, and more flexible distance computations using polygonal methods. Map projections and polygonal methods both require caution for computing distances. The former *always* distorts distances and can influence statistical estimation as we discuss later. The polygonal methods treat "distances" informally (e.g., actual roadway distance or rail track distance), rather than purely geometric concepts. Such intersite distance matrices can be imported from GIS, but they need not be valid arguments for statistical correlation functions leading to unstable or even infeasible numerical algorithms. Here, we focus upon direct distance computations and do not discuss the polygonal methods further.

We also restrict attention to point-referenced or geostatistical data where the sites are fixed, as opposed to point processes where the sites (and hence intersite distances) are random. The remainder of this article evolves by reviewing a basic framework for spatial modeling, concentrating upon isotropic models, where distances are particularly helpful for interpreting the spatial range. In Section 3, we discuss distance computations using the spherical coordinate system and map projections. Section 4 illustrates the impact of the different metrics on statistical modeling and Section 5 concludes the article with a summary.

2. Review of Spatial Regression Models

There is a growing literature on statistical modeling for point-referenced or geostatistical data. The most common setting assumes a response or a dependent variable $Y(\mathbf{s})$ observed at a generic location $\mathbf{s}$, referenced by its latitude and longitude, along with a vector of covariates $\mathbf{x}(\mathbf{s})$. One seeks to model the dependent variable in a spatial regression setting such as

$$Y(\mathbf{s}) = \mathbf{x}^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{1}$$

The residual is partitioned into a spatial process, $w(\mathbf{s})$, capturing residual spatial association, and an independent process, $\epsilon(\mathbf{s})$, also known as the *nugget* effect, modeling pure error. Inferential goals include estimation of regression coefficients, spatial and nugget variances, and the strength of spatial association through distances.

When we have observations, $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$, from n locations, we treat the data as a partial realization of a spatial process, modeled through $w(\mathbf{s})$. Hence, $w(\mathbf{s}) \sim GP(0, \sigma^2\rho(\cdot, \boldsymbol{\eta}))$ is a zero-centered Gaussian process with variance σ^2 and a valid correlation function $\rho(\cdot, \boldsymbol{\eta})$, which depends upon intersite distances ($d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$) and parameters $\boldsymbol{\eta}$ quantifying correlation decay and smoothness of process realizations. Also, we assume $\epsilon(\mathbf{s})$ are i.i.d. $N(0, \tau^2)$. Likelihood-based inference proceeds from the distribution of the data, $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma)$, with $\Sigma = \sigma^2 H(\boldsymbol{\eta}) + \tau^2 I$, where X is the matrix of covariates (or model matrix) and $H(\boldsymbol{\eta})$ is the spatial correlation matrix (corresponding to $w(\mathbf{s})$) with $H_{ij} = \rho(d_{ij}, \boldsymbol{\eta})$. See Cressie (1993) for details, including maximum likelihood and restricted maximum likelihood methods.

Statistical prediction (kriging) at a new location $\mathbf{s}_0$ proceeds from the conditional distribution of $Y(\mathbf{s}_0)$ given the data $\mathbf{Y}$ (for details see, e.g., Banerjee, Carlin, and Gelfand,

2004, p. 48–52). Collecting all the model parameters into $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\eta})$, we note that

$$\mathrm{E}[Y(\mathbf{s}_0) \,|\, \mathbf{Y}] = \mathbf{x}(\mathbf{s}_0)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\gamma}^{\mathrm{T}}\Sigma^{-1}(\mathbf{Y} - X\boldsymbol{\beta}),$$

$$\mathrm{var}[Y(\mathbf{s}_0) \,|\, \mathbf{Y}] = \sigma^2 + \tau^2 - \boldsymbol{\gamma}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma} = (\sigma^2\rho(d_{01}; \boldsymbol{\eta}), \ldots, \sigma^2\rho(d_{0n}; \boldsymbol{\eta}))$ and $d_{0j} = \|\mathbf{s}_0 - \mathbf{s}_j\|$ when $\mathbf{s}_0 \neq \mathbf{s}_j$ for all j, while the nugget effect is added to the jth entry if $\mathbf{s}_0 = \mathbf{s}_j$ for some j. Classical prediction computes the best linear unbiased predictor (BLUP) by substituting maximum likelihood estimates for the above parameters. A Bayesian solution first computes a posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{Y}) \propto f(\mathbf{Y} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})$, where $f(\mathbf{Y} \,|\, \boldsymbol{\theta})$ is the normal data likelihood and $p(\boldsymbol{\theta})$ is the prior distribution for the parameters, and then computes the posterior predictive distribution $p(Y(\mathbf{s}_0) \,|\, \mathbf{Y})$ by marginalizing over the posterior distribution, $\int f(Y(\mathbf{s}_0) \,|\, \mathbf{Y}, \boldsymbol{\theta})p(\boldsymbol{\theta} \,|\, \mathbf{Y})$.

The function $\rho(d, \boldsymbol{\eta})$ depends upon the metric used to compute d and must ensure that $H(\boldsymbol{\eta})$ is positive definite. Valid classes of correlation functions for Euclidean spaces are generated by Bochner's theorem (see, e.g., Stein, 1999), highlighting the theoretical importance of Euclidean metrics. Apparently, the parameters most sensitive to the choice of the metric are those associated with the correlation function. Also, note that the correlation function features prominently in both the likelihood as well as the predictive distribution suggesting concern regarding inferential sensitivity to the metric.

Focusing upon the correlation function parameters, we consider the flexible Matérn family, where $\boldsymbol{\eta} = (\psi, \nu)$ involves a *smoothness* parameter ν in addition to the correlation decay parameter ψ, and is given by

$$\rho(d, \nu, \psi) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}\, d\psi)^{\nu} K_{\nu}(2\sqrt{\nu}\, d\psi),$$

where $\Gamma(\,)$ is the usual Gamma function and K_{ν} is the modified Bessel function of the second kind of order ν (see, e.g., Abramowitz and Stegun, 1965). In particular, with $\nu = 0.5$, we obtain the exponential correlation function $\rho(d, \psi) = \exp(-\psi d)$. Recent interest in the study of smoothness of a spatial process and spatial gradients (Stein, 1999; Banerjee, Gelfand, and Sirmans, 2003) warrants estimation of ν. In our context it is unclear how the distance metrics will affect smoothness, so we investigate the exponential (with fixed ν) and the more general Matérn family with unknown ν.

A Bayesian framework is convenient here, allowing inference by assigning proper and moderately informative priors on the weakly identified correlation function parameters. For example, for the smoothness parameter ν in the Matérn, we can follow Stein (1999) that the data cannot distinguish between $\nu = 2$ and $\nu > 2$, which suggests placing a *Unif*$(0, 2)$ prior on ν. Usually a Markov chain Monte Carlo (MCMC) algorithm is required to obtain the joint posterior distribution of the parameters, but again there are different strategies to opt for. For example, we may work with the marginalized likelihood as above, $\mathbf{Y} \,|\, \boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\boldsymbol{\eta}) + \tau^2 I)$, or we may add a hierarchy with spatial random effects $\mathbf{W} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))$ such that

$$\mathbf{Y} \,|\, \boldsymbol{\theta}, \mathbf{W} \sim N(X\boldsymbol{\beta} + \mathbf{W}, \tau^2 I),$$

$$\mathbf{W} \sim N(\mathbf{0}, \sigma^2 H(\boldsymbol{\eta})).$$

In either framework, a Gibbs sampler may be designed, with embedded Metropolis or slice-sampling steps, to obtain the marginal posterior distribution (see, e.g., Banerjee et al., 2004).

Much more complex hierarchical models have been discussed extensively in the spatial literature but, irrespective of their complexity, they typically incorporate a spatial correlation function whose computation involves intersite distance computations. Therefore, although we work with simpler isotropic spatial models, our results will be relevant in a broader context.

3. Computing Distances

We consider a few different approaches for computing distances on the Earth, classifying them as those arising from the classical spherical coordinates, and those arising from planar projections. Our treatment is comparative, eliciting some nontrivial aspects that impact spatial modeling. While spherical geometry may suggest natural metrics (such as the geodetic metric to be discussed shortly), we do not recommend a *true* distance metric because none may be appropriate for scientific data analysis. Henceforth, $\|\cdot\|$ will denote the Euclidean metric in $\Re^2$ or $\Re^3$ as the case may be.

Recall that a spherical model of the Earth is divided by parallels of latitude, referencing east–west, and the meridians of longitude that are great-circle arcs (circle passing through the two points with center as the center of the Earth) joining the poles, intersecting the parallels orthogonally. The Earth is not exactly a sphere, but an ellipsoid (surface obtained by revolving an ellipse). For geodetic computations requiring very high degrees of accuracy, the ellipsoidal model of the Earth is used, but for spatial modeling a spherical model suffices. In fact, apart from locations in the polar regions the accuracy of the spherical model is excellent.

3.1 *Spherical Coordinates, the Geodetic Formula, and Euclidean Approximations*

Figure 1 shows the spherical coordinate system, where $P_1 = (\lambda_1, \theta_1)$ and $P_2 = (\lambda_2, \theta_2)$ are two points on the surface of the Earth (sphere not shown) with center O, given by longitudes λ_1 and λ_2, and latitudes θ_1 and θ_2. The geodetic distance is the length of the arc of a great circle joining P_1 and P_2 and is obtained as $R\phi$, where R is the radius of the Earth and ϕ is the angle between the vectors OP_1 and OP_2. A three-dimensional orthogonal coordinate system (x, y, z) is set up with the origin at the center O, the z-axis directed toward the North Pole, the x-axis, on the equatorial plane, along the Greenwich meridian (the $0°$ meridian, passing through Greenwich England), and the y-axis perpendicular to the x-axis on the equatorial plane.

Using projections P_1' and P_2' on the x–y plane, we obtain

$$(x, y, z) = (R\cos\theta\cos\lambda, R\cos\theta\sin\lambda, R\sin\theta).$$

Letting $\mathbf{u}_1 = (x_1, y_1, z_1)$ and $\mathbf{u}_2 = (x_2, y_2, z_2)$ be the unit vectors $\overrightarrow{OP_1}$ and $\overrightarrow{OP_2}$, our desired angle ϕ is given by $\cos\phi = \langle \mathbf{u}_1, \mathbf{u}_2 \rangle$, where $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle$ denotes the inner product between these vectors. Simple trigonometric identities reveal the geodetic distance as

$$R\phi = R \arccos(\sin\theta_1\sin\theta_2 + \cos\theta_1\cos\theta_2\cos(\lambda_2 - \lambda_1)). \quad (2)$$

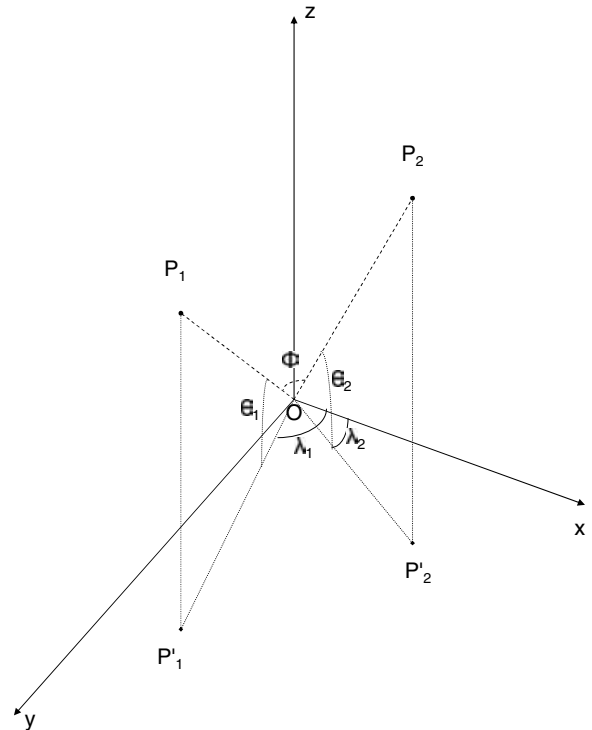**Figure 1.** The spherical coordinate system. The shortest distance from P_1 to P_2 is through the great circle arc (not shown) formed by intersecting the sphere with the plane containing O, P_1, and P_2.

This is given in Cressie (1993, p. 265) as the great-arc distance. The correct scale is obtained with the angle ϕ in radian measure with the distance expressible in kilometers or miles depending upon the unit of R. Using $R = 6371$ km results in a sufficiently good approximation. For example, to obtain the geodetic distance between Chicago (87.63W, 41.88N) and Minneapolis (93.22W, 44.89N), we plug in the appropriate values in (2) to obtain the required distance as $6371 \times \arccos(0.9961) \approx 562$ km.

The transcendental nature of equation (2) dispels any misconception that the relationship between the Euclidean distances and the geodetic distances is just a matter of scaling and merits further investigation. A simple scaling of the geographical coordinates results in a "naive Euclidean" metric (as is done by Kaluzny et al., 1998 and Ecker and Gelfand, 1997) obtained directly in degree units, and converted to kilometer units as $\|P_1 - P_2\|\pi R/180$. This metric performs well on small domains but always overestimates the geodetic distance, *flattening out* the meridians and parallels, and stretching the curved domain onto a plane, thereby stretching distances as well. As the domain increases, the estimation deteriorates.

A more natural metric to consider is along the "chord" joining the two points. This is simply the Euclidean metric $\|\mathbf{u}_2 - \mathbf{u}_1\|$, yielding a "burrowed through the Earth" distance—the chordal length between P_1 and P_2. The slight underestimation of the geodetic distance is expected, because the chord "penetrates" the domain, producing a straight line approximation to the geodetic arc.

Table 1
Comparison of different methods of computing distances

Methods	Colorado data (farthest)	Chicago–Minneapolis	New York–New Orleans
Geodetic	741.7 km	562.0 km	1897.2 km
Naive Euclidean	933.8 km	706.0 km	2172.4 km
Chord	741.3 km	561.8 km	1890.2 km
Mercator	951.8 km	773.7 km	2336.5 km
Sinusoidal	742.7 km	562.1 km	1897.7 km
Centroid-based	738.7 km	562.2 km	1901.5 km

The first three rows of Table 1 compare the geodetic distance with the "naive Euclidean" and chordal metrics. The first column corresponds to the distance between the farthest points in a spatially referenced data set comprising 50 locations in Colorado (more of this in Section 4), while the next two present results for two differently spaced pairs of cities. The overestimation and underestimation of the "naive Euclidean" and "chordal" metrics, respectively, is clear, although the chordal metric excels even for distances over 2000 km (New York and New Orleans).

This approximation of the chordal metric has an important theoretical implication for the spatial modeler. A troublesome aspect of geodetic distances is that they are *not* necessarily valid arguments for correlation functions defined on Euclidean spaces (e.g., the exponential, spherical, Matérn, etc.). However, the excellent approximation of the chordal metric (which is Euclidean) ensures that in most practical settings, as in our illustration in Section 4, valid correlation functions in $\Re^3$ such as the Matérn and exponential yield positive definite correlation matrices with geodetic distances and enable proper convergence of the statistical estimation algorithms.

Schoenberg (1942) develops a necessary, sufficient representation for valid positive-definite functions on spheres in terms of normalized Legendre polynomials P_k of the form:

$$\psi(t) = \sum_{k=0}^{\infty} a_k P_k(\cos t),$$

where a_k's are positive constants such that $\sum_{k=0}^{\infty} a_k$ converges. An example is given by

$$\psi(t) = \frac{1}{\sqrt{1 + \alpha^2 - 2\alpha \cos t}}, \quad \alpha \in (0, 1),$$

which can be easily shown to have the Legendre polynomial expansion $\sum_{k=0}^{\infty} \alpha^k P_k(\cos t)$.

The chordal metric also provides a simpler way to construct valid correlation functions over the sphere using a sinusoidal composition of any valid correlation function on Euclidean space. To see this, consider a unit sphere ($R = 1$) and note that

$$\|\mathbf{u}_1 - \mathbf{u}_2\| = \sqrt{2 - 2\langle \mathbf{u}_1, \mathbf{u}_2 \rangle} = 2\sin(\phi/2).$$

Therefore, a correlation function $\rho(d)$ (suppressing the range and smoothness parameters) on the Euclidean space transforms to $\rho(2\sin(\phi/2))$ on the sphere, thereby *inducing* a valid correlation function on the sphere. This has several advantages over the Legendre polynomial approach of Schoenberg: (1) We retain the interpretation of the smoothness and decay parameters, (2) is simpler to construct and compute, and

(3) builds upon a rich legacy of investigations (both theoretical and practical) of correlation functions on Euclidean spaces. We do not explore spherical correlation functions here, restricting ourselves to the Matérn, and its special case, the exponential, correlation functions that are popular in practice.

3.2 *Map Projections*

An alternative approach to using Euclidean metrics is that of a planar projection of the spatial domain. This is particularly popular among GIS users, where several map projections are available, and has the added advantage of working with two-dimensional coordinates, unlike the three-dimensional chordal metric. In fact, currently most existing spatial statistics software (e.g., `WinBUGS` [Thomas et al., 2002], `geoR` [Ribeiro and Diggle, 2003]) allow specification of only two-dimensional Euclidean coordinates.

We will restrict ourselves to the purely mathematical map projections that derive a relationship between geographical coordinates (λ, θ) and cartesian coordinates (x, y) through

$$x = f(\lambda, \theta), \quad y = g(\lambda, \theta),$$

where f and g are functions that are determined by mapping infinitesimal quadrilaterals with desirable map properties. Ideally, we would seek to preserve all intersite distances but the existence of such a projection is precluded by Gauss' Theorema Eggregium in differential geometry (see, e.g., Guggenheimer, 1977, p. 240–242). Projections such as the gnomonic projection (Snyder, 1987, p. 164–168) give the correct distance from a single reference point, but is less useful for the practicing spatial analyst who needs to obtain complete intersite distance matrices, which would require, not one, but several such maps.

Areas and angles can, however, be preserved and most mathematical projections offered by GIS can be classified as either *conformal* (preserving angles) or *equal-area* (preserving areas). These are developed using geometric constructions or differential geometric analysis to provide underdetermined systems of partial differential equations with further map properties leading to the final equations. See Banerjee et al. (2004, p. 12–14) for some heuristic derivations. Distances are *always* distorted, with its extent varying by type, but typically conformal projections distort distances much more than equal-area.

We illustrate with two popular projections, one of each type: the Mercator (conformal) and the sinusoidal (equal-area). The Mercator projection is a classical conformal projection where *loxodromes* (curves that intersect the meridians at a constant angle) are straight lines on the map, a property particularly useful for navigation purposes, derived by letting

$\partial g / \partial \theta = R \sec \theta$. After suitable integration, this leads to the analytical equations (with the 0° meridian as the central meridian),

$$f(\lambda, \phi) = R\lambda, \quad g(\lambda, \phi) = R \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right). \quad (3)$$

The sinusoidal projection yields equally spaced rectilinear parallels (with the 0° meridian as the central meridian), by specifying

$$f(\lambda, \theta) = R\lambda \cos \theta, \quad g(\lambda, \theta) = R\theta. \quad (4)$$

These and several other projections are routinely available in the GIS software, and in interfaces such as the R package `mapproj` (McIlroy, 2004), but are simple enough to be computed without accessing such packages.

Yet another class of projections are *site-adaptive* in that they use information on the specific configuration of the sites (the data). One such projection, which we call *centroid-based*, sets up rectangular axes along the *centroid* of the observed locations and scales the points according to these axes. Thus, with N locations $(\lambda_i, \theta_i)_{i=1}^N$, we first compute the centroid $(\bar{\lambda}, \bar{\theta})$ (the mean longitude and latitude). Next, two geodetic distances are computed that scale the axes: d_X is the geodetic distance (using (2)) between $(\bar{\lambda}, \theta_{\min})$ and $(\bar{\lambda}, \theta_{\max})$, where $\theta_{\min}$ and $\theta_{\max}$ are the minimum and maximum of the observed latitudes; analogously, d_Y is that between $(\lambda_{\min}, \bar{\theta})$ and $(\lambda_{\max}, \bar{\theta})$. This centroid-based projection then defines a two-dimensional, planar coordinate system as scaled displacements with respect to the axes along the centroid:

$$x = \frac{\lambda - \bar{\lambda}}{\lambda_{\max} - \lambda_{\min}} d_X, \quad y = \frac{\theta - \bar{\theta}}{\theta_{\max} - \theta_{\min}} d_Y. \quad (5)$$

Returning to the bottom half of Table 1, we compare the three projections in equations (3)–(5). We find that the sinusoidal and centroid-based projections seem to be distorting distances much less than the Mercator, which performs even worse than the naive Euclidean. Their impact upon statistical estimation and prediction will be discussed in Section 4.

Note that Table 1 is more pertinent from a geographical or geodetic viewpoint than for the spatial statistician, as they do not reflect how statistical estimation is affected, where points that are "closer" together have greater influence on analysis. Nevertheless, the distortion brought about by a poor metric alters the definition of "closeness" and can lead to erroneous statistical estimates (see Section 4).

The centroid-based projection has the potentially unattractive property of being data dependent in that its computation changes with new sites being added. Addition of new sites is particularly common in spatiotemporal settings such as environmental monitoring, and (5) needs to be recomputed every time. Because the sinusoidal does not suffer from this, has comparable accuracy, and is easy to compute, it might be preferred. Nevertheless, being site-adaptive the centroid-based projection is more flexible than the sinusoidal and *may* perform better for certain configurations. Also, it is inexpensive to compute and, unless the number of sites is huge, presents itself as a viable alternative.

We conclude this section with a brief discussion of the Universal Transverse Mercator (UTM) projection system. Rather than a purely mathematical projection, the UTM is more of a *coordinate* or *grid* system (also known as state plane coordinates) using a *transverse* aspect of the Mercator projection (see, e.g., Snyder, 1987). The projection equations are further transformed into "Easting–Northing" coordinates by overlaying a grid that divides the domain into zones each 6° wide, referencing each point from a zone-specific central meridian. While these UTM grids can be used to adjust for local scale to provide accurate measurements, they are in the same scale as the chordal or the sinusoidal. However, these accrue additional computational complexity (for the grid) and should always be imported from the GIS software or interfaces; yet many GIS interfaces do not provide them (e.g., `mapproj`). For these reasons, we do not explicitly use them in this article, although their use typically produces accurate results comparable to the geodetic metric.

4. Illustration

We illustrate spatial modeling under different geodetic computations with a weather data set obtained from the National Center for Atmospheric Research (NCAR), Boulder, Colorado with the mean temperature measurements (in 10° C units) obtained at 50 sites, in the month of January in 1997 as our dependent variable $Y(\mathbf{s})$. Also supplied is the elevation (in 100 m units) at each site, so the covariate $\mathbf{x}(\mathbf{s})$ comprises an intercept and elevation. A univariate spatial model as in (1) explains temperature given elevation, accounting for the spatial correlation in the data. Figure 2 shows an elevation map of the spatial domain with the solid circles indicating our sampling locations. The contours represent a particularly interesting topography where temperature is expected to show rich spatial variation. A detailed spatiotemporal analysis of this data set using coregionalized models is performed in Gelfand, Banerjee, and Gamerman (2005).
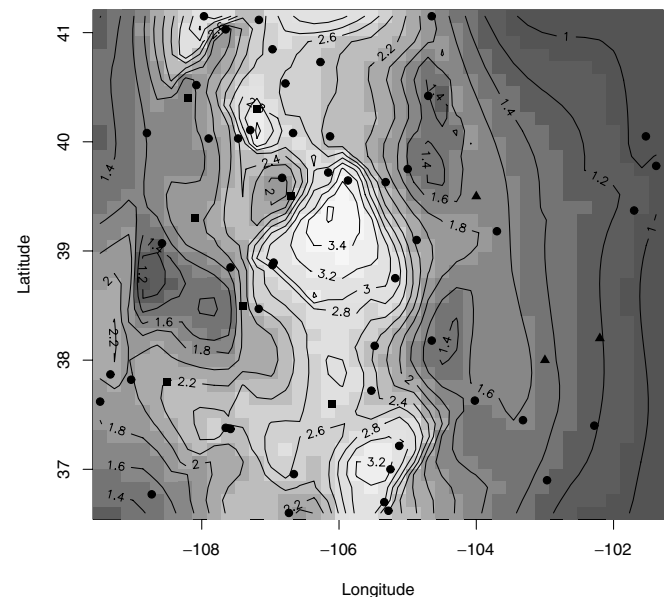

Figure 2. An image plot of the spatial domain in Colorado, with elevation contours. The 50 sampling sites are indicated by the solid circles. Three solid triangles represent the three sensitive prediction locations while seven solid squares represent the remaining prediction locations seen in Figure 3.

Table 2
Parameter estimates (medians and 95% posterior credible intervals) for the exponential correlation model under different metrics

Parameter	Geodetic	Naive Euclidean	Chordal
Intercept	1.031 (0.501, 1.518)	1.020 (0.321, 1.607)	1.141 (0.750, 1.488)
Elevation	-0.417 (-0.530, -0.300)	-0.428 (-0.528, -0.331)	-0.419 (-0.512, -0.315)
σ^2	0.098 (0.031, 0.231)	0.110 (0.033, 0.341)	0.128 (0.035, 0.483)
ψ	1.09×10^{-2} (0.69×10^{-2}, 7.63×10^{-2})	0.71×10^{-2} (0.27×10^{-2}, 5.42×10^{-2})	1.12×10^{-2} (0.74×10^{-2}, 8.50×10^{-2})
Range	275.2 (39.3, 434.8)	422.5 (55.4, 1109.2)	267.8 (35.3, 405.4)
τ^2	0.011 (0.003, 0.017)	0.011 (0.004, 0.024)	0.008 (0.003, 0.024)

Parameter	Mercator	Sinusoidal	Centroid-based
Intercept	1.015 (0.307, 1.551)	1.035 (0.399, 1.569)	1.103 (0.601, 1.643)
Elevation	-0.430 (-0.532, -0.327)	-0.432 (-0.533, -0.329)	-0.426 (-0.521, -0.321)
σ^2	0.109 (0.031, 0.351)	0.105 (0.030, 0.348)	0.093 (0.033, 0.622)
ψ	0.66×10^{-2} (0.19×10^{-2}, 5.24×10^{-2})	1.12×10^{-2} (0.71×10^{-2}, 8.49×10^{-2})	1.15×10^{-2} (0.71×10^{-2}, 7.80×10^{-2})
Range	454.5 (57.25, 1578.9)	267.9 (35.3, 422.5)	260.9 (38.5, 422.5)
τ^2	0.011 (0.003, 0.025)	0.011 (0.004, 0.023)	0.010 (0.003, 0.018)

We apply each of the six metrics in Table 1 to the exponential and Matérn functions. We also analyzed the data using UTM projections for which our results were almost identical to that of the geodetic metric and, hence, are not presented. We performed classical likelihood-based as well as Bayesian analysis for the exponential models, but only a Bayesian analysis for the Matérn (see Section 2). Because the classical and Bayesian methods provided extremely consistent answers for the exponential, we present results only from the latter.

We adopted a flat prior for β (the intercept coefficient), and relatively vague $IG(0.001, 0.001)$ (Inverted-Gamma) priors for σ^2 and τ^2. We also chose a Gamma prior for the correlation decay parameter η, specified so that the prior spatial range has a mean of about half of the maximum intersite distance in our data, obtained from the first column of Table 1 for the respective metric. Practical analysis calculates the spatial range by solving for $\rho(d; \eta) = 0.05$. In addition, for the Matérn correlation function we use a $U(0, 2)$ prior for the smoothness parameter in our data.

Three parallel MCMC chains were run for 10,000 iterations. The CODA package in R was used to diagnose convergence by monitoring mixing, Gelman–Rubin diagnostics, autocorrelations, and cross-correlations. In each case, 5000 iterations were enough for sufficient mixing of the chains, so the remaining 15,000 samples (5000 × 3) were retained for posterior analysis. We used C/C++ code to fit these models with posterior summarization in R. We remark that implementations with the naive Euclidean and projection methods with an exponential correlation function could be performed in WinBUGS and geoR, because they need a *two-dimensional* coordinate input. The Matérn is accessible only in the latter, but with a fixed smoothness parameter. Classical analysis for the same could also be performed in geoR (see, e.g., Banerjee et al., 2004, p. 64–65).

Tables 2 and 3 show the parameter estimates (medians) with 95% credible intervals for the exponential and Matérn correlation functions, respectively, under different choices of the distance metric. We see that the regression estimates are

Table 3
Parameter estimates (medians and 95% posterior credible intervals) for the Matérn correlation model under different metrics

Parameter	Geodetic	Naive Euclidean	Chordal
Intercept	1.087 (0.789, 1.410)	1.031 (0.664, 1.447)	1.015 (0.707, 1.308)
Elevation	-0.430 (-0.533, -0.336)	-0.421 (-0.523, -0.322)	-0.422 (-0.515, -0.329)
σ^2	0.171 (0.043, 1.539)	0.097 (0.033, 0.505)	0.093 (0.034, 0.435)
ψ	7.47×10^{-3} (4.79×10^{-3}, 51.18×10^{-3})	4.26×10^{-3} (2.27×10^{-3}, 41.42×10^{-3})	7.63×10^{-3} (4.85×10^{-3}, 54.51×10^{-3})
ν	0.770 (0.213, 1.413)	0.819 (0.227, 1.426)	0.742 (0.199, 1.402)
Range	273.7 (39.1, 426.7)	477.3 (48.2, 895.6)	268.7 (36.8, 422.8)
τ^2	0.008 (0.003, 0.019)	0.007 (0.003, 0.018)	0.008 (0.003, 0.017)

Parameter	Mercator	Sinusoidal	Centroid-based
Intercept	1.015 (0.332, 1.597)	1.014 (0.377, 1.603)	1.088 (0.735, 1.511)
Elevation	-0.426 (-0.527, -0.333)	-0.431 (-0.530, -0.330)	-0.427 (-0.527, -0.324)
σ^2	0.111 (0.033, 0.363)	0.106 (0.031, 0.321)	0.102 (0.035, 0.684)
ψ	4.01×10^{-3} (1.98×10^{-3}, 40.01×10^{-3})	7.61×10^{-3} (4.89×10^{-3}, 55.08×10^{-3})	7.57×10^{-3} (4.74×10^{-3}, 53.22×10^{-3})
ν	0.843 (0.331, 1.533)	0.767 (0.225, 1.441)	0.791 (0.212, 1.421)
Range	506.9 (51.7, 1025.5)	270.3 (38.9, 418.0)	269.5 (37.4, 430.2)
τ^2	0.011 (0.004, 0.024)	0.009 (0.004, 0.024)	0.007 (0.003, 0.017)

virtually unaffected by the metric; in each case there is a significantly positive intercept and, quite expectedly, a significantly negative effect of elevation on temperature. The spatial variance σ^2 seems to explain a substantial portion of the residual variation, dominating the nugget effect τ^2. For example, in the geodetic setting with exponential correlation functions the spatial variance explains about $\sigma^2/(\sigma^2 + \tau^2) \approx$ 90% of spatial variation, while with the Matérn function this is about 96%. This seems to be quite stable across the different metrics. For the Matérn correlation function, the smoothness parameter ν also seems to be robustly estimated across metrics, with 0.5 included in each of the intervals, but the median
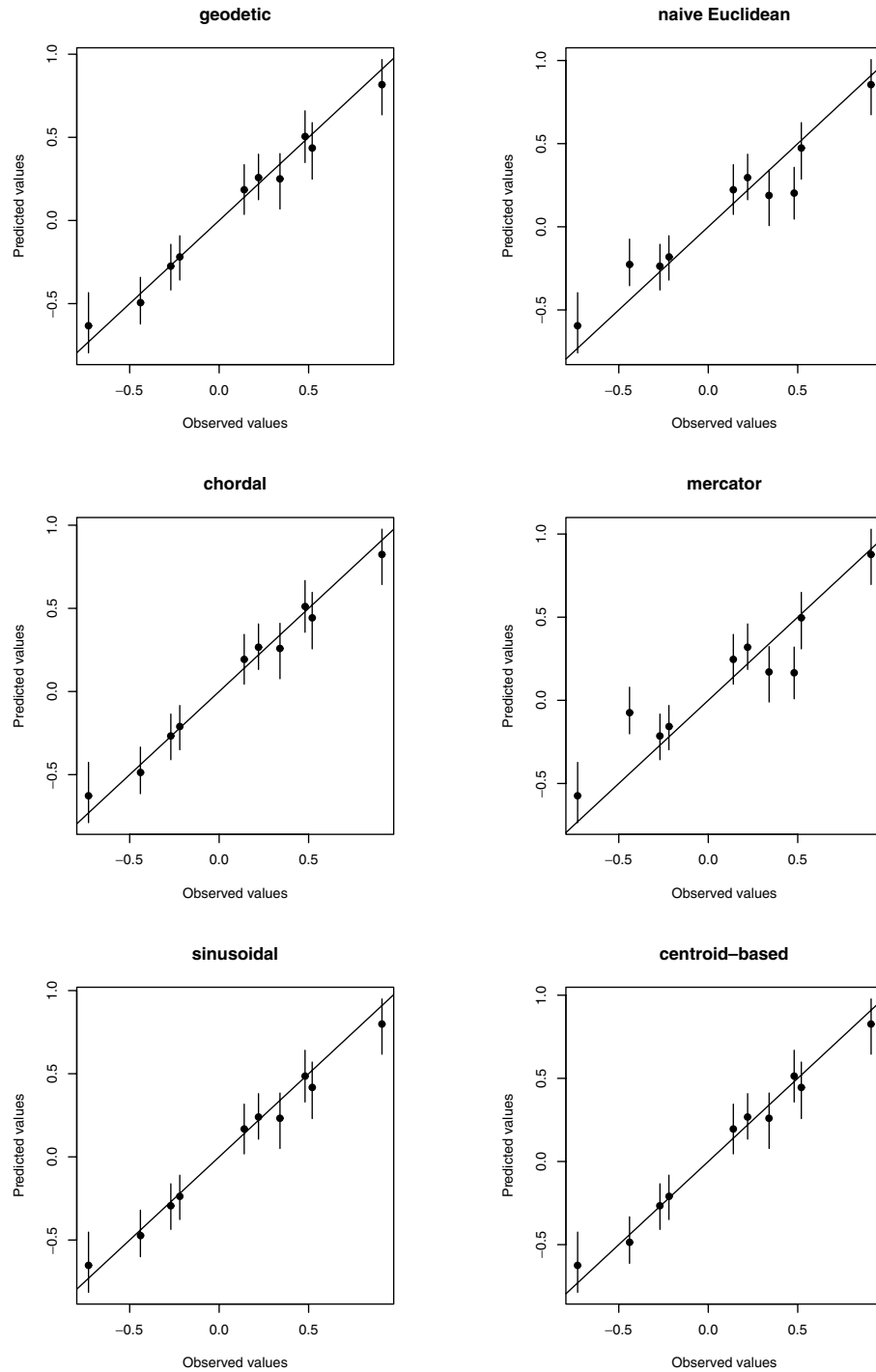

Figure 3. The predictive performance of the model under the six different metrics for the Matérn model. The results for the exponential model are almost identical and not shown.

seems to shift to slightly higher values for the Mercator and naive Euclidean indicating slight oversmoothing compared to the other four.

The estimate that seems to be most sensitive to the choice of the metric is ϕ, the correlation parameter, and hence the implied spatial range. While on the one hand this is to be expected, being in some sense "closest" to the distance metric, this effect is interesting because larger distances (where these metrics really differ) are downweighted by the correlation functions. In fact, we see that the estimated spatial range using the chordal, sinusoidal, and centroid-based metrics are similar to the range estimated from the geodetic metric. In comparison, the naive Euclidean metric and Mercator's projection estimate the spatial range by a factor exceeding 1.5 times the geodetic range for the exponential correlation function; this is even more drastic for the Matérn. Apparently this discrepancy seems to be consistent with the purely geographical effects in Table 1, resulting from a spurious expansion of the spatial domain. In the same vein, a benign difference (lower range than the geodetic) is seen with the chordal approximation, not surprising given the minor bias inherent in its definition. The estimates from the sinusoidal and centroid-based projections are also quite close to the geodetic, corroborating their claim as viable alternatives.

Turning next to predictive performance of these models, we assess the models for 10 holdout locations (with known elevation and temperature) and predict using our fitted model. Figure 3 plots the predicted values against the observed values with the dots representing the predictive mean and the bars representing 95% prediction intervals. Because there is no inherent ordering of the sites, we sort them by the (*true*) observed values arranged along the $y = x$ line to better display discrepancies. While the geodetic, chordal, sinusoidal, and centroid-based projections all seem to predict well for all these sites, three sites (the second, seventh, and eighth in ascending order in each of the panels in Figure 3) seem to be quite sensitive to the choice of the metric with their prediction intervals not including the observed value (along the $y = x$ line). These three sites are indicated by solid triangles in Figure 2 and are relatively isolated (in the eastern part) from the sampling sites, while the remaining seven holdout locations indicated by solid squares in Figure 2 are located amidst the sampling locations and show more robust predictive performance. The results for the exponential correlation model are almost identical and not shown.

Finally, we computed the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) for our models to investigate how the model choice criterion captures metric discrepancies. Briefly, we summarize the *fit* of the model using the posterior expectation $\bar{D} = E_{[\boldsymbol{\theta}|\mathbf{y}]}[D(\boldsymbol{\theta})]$, where $D(\boldsymbol{\theta})$ is the *deviance* statistic $-2\log f(\mathbf{y}\,|\,\boldsymbol{\theta})$. The model is penalized by the effective number of parameters p_D estimated as $\bar{D} - D(E[\boldsymbol{\theta}\,|\,\mathbf{y}])$. The DIC is then computed as the sum of $\bar{D}$ and p_D.

Table 4 shows these computations for the exponential and Matérn models under the six different metrics. Generally, the Matérn seems to perform better for this data set for each of the metrics, even for the naive Euclidean metric where its overestimation (relative to the geodetic) of the spatial range is more drastic than the exponential. Apparently, in spite

Table 4
Deviance information criterion (*DIC*) *for model choice*

Methods	Exponential			Matérn		
	p_D	$\bar{D}$	DIC	p_D	$\bar{D}$	DIC
Geodetic	7.25	11.83	18.08	7.23	9.80	17.03
Naive Euclidean	9.52	13.15	22.67	9.11	12.28	21.39
Chordal	7.18	11.92	19.10	7.29	9.86	17.15
Mercator	10.21	14.65	24.86	10.14	13.71	23.85
Sinusoidal	7.23	11.86	19.09	7.31	10.01	17.32
Centroid-based	7.57	11.98	19.55	7.41	10.28	17.69

of this overestimation, the Matérn's flexibility in capturing process smoothness (see, e.g., Stein, 1999) leads to a better fit than the exponential. In fact, the criterion also seems to capture the difference between the metrics with the Mercator and the naive Euclidean metric having higher scores and the other four performing much better.

5. Discussion

This article explored different options for approximating geodetic distances providing theoretical clarifications and investigating impact upon statistical modeling and focusing upon ease of implementation. While geometrically natural metrics are not necessarily the most appropriate, often it is difficult to ascertain what is. Keeping this in mind, the article has tried to convey that (1) the practicing spatial analyst is often confronted with a choice of metrics which is not easily resolvable with scientific information, and (2) the choice of the metric *may* influence geostatistical analysis including estimation and prediction of certain parameters. We have demonstrated that uninformed formulation of metrics can affect estimation of the spatial range leading to differences in predictive performance. Viable solutions have been proposed and some have been shown to work well. Effects of spurious nonstationarity and anisotropy can be undertaken as further investigations. Future work can also focus upon the point-process settings, where intersite distances arise as *random* processes, with sensitivity of spatial tests of randomness on choice of metric.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. New York: Dover.

Banerjee, S., Gelfand, A. E., and Sirmans, C. F. (2003). Directional rates of change under spatial process models. *Journal of the American Statistical Association* **98,** 946–954.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman and Hall/CRC Press.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd edition. New York: Wiley.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics* **2,** 347–369.

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, in press.

Guggenheimer, H. W. (1977). *Differential Geometry*. New York: Dover.

Jones, C. B. (1997). *Geographical Information Systems and Computer Cartography*. Harlow, Essex, U.K.: Addison Wesley Longman.

Kaluzny, S. P., Vega, S. C., Cardoso, T. P., and Shelly, A. A. (1998). *S+ Spatial Stats*. New York: Springer.

McIllroy, D. (2004). *The Mapproj Package.* Available at `http://cran.r-project.org`.

Ribeiro, P. J. and Diggle, P. J. (2003). *The GeoR Package.* Available at `http://www.est.ufpr.br/geoR`.

Schoenberg, I. J. (1942). Positive definite functions on spheres. *Duke Mathematics Journal* **9,** 96–108.

Snyder, J. P. (1987). *Map projections—A working manual.* United States Geological Survey Professional Paper 1395.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B* **64,** 583–639.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging.* New York: Springer.

Thomas, A., Best, N., Arnold, R., and Spiegelhalter, D. (2002). *The GeoBUGS User Manual.* Available at `http://www.mrc-bsu.cam.ac.uk/bugs`.