The Canadian Journal of Statistics
Vol. 34, No. 4, 2006, Pages 659–676
La revue canadienne de statistique

Modelling map positional error to infer true feature location

Jarrett J. BARBER, Alan E. GELFAND and John A. SILANDER, Jr.

Key words and phrases: Bayesian inference; Bayesian model averaging; Berkson model; bivariate spatial process; coregionalization; measurement error.

MSC 2000: Primary 62F15, 62M30.

Abstract: The authors consider the issue of map positional error, or the difference between location as represented in a spatial database (i.e., a map) and the corresponding unobservable true location. They propose a fully model-based approach that incorporates aspects of the map registration process commonly performed by users of geographic informations systems, including rubber-sheeting. They explain how estimates of positional error can be obtained, hence estimates of true location. They show that with multiple maps of varying accuracy along with ground truthing data, suitable model averaging offers a strategy for using all of the maps to learn about true location.

Modélisation de l'erreur associée à la position d'un objet sur une carte en vue de sa localisation

Résumé : Les auteurs s'intéressent à l'erreur associée à la position d'un objet sur une carte, c'est-à-dire à la différence entre sa position telle que représentée par ses coordonnées spatiales (c'est-à-dire sur une carte) et sa véritable localisation inconnue dans l'espace. Ils proposent un modèle tenant compte de différents aspects des procédés de cartographie tels que mis en œuvre dans les systèmes d'information à référence géographique, y compris la correction géométrique par membrane élastique. Ils expliquent comment estimer l'erreur associée au relevé et donc la position réelle d'un objet. Ils montrent qu'en opérant une moyenne sur différents modèles et en s'aidant de cartes de précision variée et de données de contrôle au sol, on peut arriver à déterminer la véritable position de l'objet.

1. INTRODUCTION

1.1. The positional error problem.

Geographic information systems (GIS) have become a popular and important way to store, manipulate, and analyse spatial data. See the two-volume introduction to GIS given by Longley, Goodchild, Maguire & Rhind (1999a, b). The development of automated mapping and GIS has resulted in an enormous amount of information stored in the form of spatial databases. Methods for assessing the accuracy of these databases are comparably underdeveloped and the apparent sophistication with which spatial data is depicted in the form of maps often conveys a sense of accuracy that may be unwarranted. In response, the GIS community has begun to devote effort to the issue of spatial database accuracy (Goodchild & Gopal 1989; Guptill & Morrison 1995).

Spatial data quality spans a broad range of topics (Goodchild & Gopal 1989; Thapa & Bossler 1992; Guptill & Morrison 1995; Veregin 1999; Lowell & Jaton 1999; Mowrer & Congalton 1999; Shi, Fisher & Goodchild 2002). The more statistically refined approaches to assessing uncertainty are based mostly on existing methods in geostatistics (Atkinson 1999) and are applied to quantitative attributes recorded at locations. However, in nearly all such methods, locations are assumed to be known, or a priori measures of positional accuracy (Thapa & Bossler 1992; Drummond 1995; Veregin 1999) are used to gain some insight into the uncertainty arising from positional error. See, however, Kiiveri (1997) for a notable exception.

The contribution of this paper is to address the issue of map positional error, or the difference between location as represented in a spatial database (i.e., map) and the corresponding,

unobservable true location, in the presence of reference location information of higher accuracy than the map locations. Notions of positional uncertainty in lines originate with the idea of the epsilon-band (Perkal 1966; Chrisman 1982; Blakemore 1984) which has been generalised by various authors who typically assume some form of bivariate Gaussian distribution for the positional error components of the points that define a line (Caspary & Scheuring 1993; Leung & Yan 1998; Shi 1998; Shi & Liu 2000).

Our data consist of feature locations on one or more maps as well as control points—so-called ground-truthing—associated with a subset of the mapped feature locations. Here, we take the control points to be Global Position System (GPS) measured locations, but, alternatively, they might be locations on our highest quality map. Regardless, they are not viewed as the true locations; there is still measurement error. Our objective is to explain positional error and to infer the true location of feature coordinates represented on one or more maps, including attaching realistic uncertainty to this inference. As we explain in Section 2, for us, *true* location is not geodetic but rather true projected coordinates relative to a specified reference frame. Thus, our contribution is focused on improved map making. For instance, bringing data from lower quality maps to higher quality maps on which they are absent is an issue of interest in the GIS community. We do not envision application of our approach to accuracy issues of concern in the surveying and geodesy communities.

1.2. The modelling problem.

The problem of positional error associated with maps has received essentially no attention in the statistical literature. That is, by now there is a large literature on modelling uncertainty associated with measurements at locations. See, e.g., the books by Cressie (1993) and by Banerjee, Carlin & Gelfand (2004). All of this work assumes that the locations are correct, i.e., that the location associated with the measurement is, in fact, the exact location where the measurement was taken. Notable exceptions are the work of Gabrosek & Cressie (2002) and the follow-on work of Cressie & Kornak (2003). These papers consider two location error models referred to by Cressie & Kornak (2003) as the coordinate positioning error model and the feature positioning model. In either case, the effort is to assess the effect on spatial prediction—point and interval estimates—without interest in making inferences about true feature location.

To clarify our setting, with regard to a given map, we have three bivariate variables to consider: a feature location on the map, an associated GPS location, and a true location for that feature. At most the first two are observed. In fact, the set of GPS measured locations is only a small subset of the feature locations on the map; rarely do we have a GPS location without an associated feature location. Evidently, the relationship between true location and GPS location is not map dependent; rather, it is a reflection of the accuracy of the GPS measuring equipment. The process of explaining GPS locations using map features is referred to as *registration* of the map. So, we can cast the map positional error problem as one of registering GPS locations against the given map followed by inferring the true location—hence positional error—based upon our knowledge of GPS accuracy.

Hence, associated with any map feature location, there is a potential GPS location that may or may not have been observed. In the latter case we have a prediction problem within the registration model for the map. Also, there is a true location associated with every map feature and this always entails a prediction problem regardless of whether an associated GPS location has been recorded. Furthermore, there is a potential *inverse* problem which may be of interest, i.e., learning about an unobserved feature location on a given map associated with an observed GPS location. This inverse problem is accessed through the registration model for the map. However, there is no inverse problem with regard to true location since the latter is never observed.

With interest in capturing uncertainty well, it is promising to formalise the modelling within a Bayesian framework. For such modelling, it is straightforward to view feature location on a map as an explanatory variable for GPS location, the latter viewed as a response. Given the large number of map feature locations compared with the small number of GPS locations, attrac-

tively this provides more prediction than inversion. It is also in agreement with the customary GIS map registration approach; see, e.g., Dowman (1999). Then, the true location, since unknown, is taken to be random, varying around the GPS location. Expressed in different terms, we have a measurement error model in the response variable described using a "Berkson" specification (Fuller 1987; Carroll, Ruppert & Stefanski 1995). Alternatively, we could view the map feature locations as driving the true locations with the GPS locations varying around the true locations, the measurement error model perspective (again, Fuller 1987; Carroll, Ruppert & Stefanski 1995). In this case, we would marginalise over the true locations, fit a registration model with increased uncertainty and then back out the true locations. We anticipate inference with regard to positional error will be very similar to that obtained using the Berkson specification.

A different modelling approach might be suggested, i.e., to view the map feature location as the response with the GPS location as the observed covariate level and the true location as the actual covariate level. This casts the problem in terms of measurement error in the covariate. Then, analogously to the preceding paragraph, we have the possibility of modelling GPS location as varying around true location (the measurement error model) or true location varying around GPS location (Berkson). Regardless, the true location would be presumed to drive map feature location; it doesn't make sense to assume that the GPS location does. Unfortunately, such a model is intractable to fit. That is, we anticipate spatial dependence in positional error; the error vector associated with a given location is expected to be more highly associated with that of a nearby location than with that of a more distant location. To capture such spatial dependence we require a bivariate spatial process model. Such a model will in turn require a valid cross-covariance function to provide association between pairs of positional errors and, indeed, with n positional vectors, will result in an $2n \times 2n$ cross covariance matrix for these n 2×1 error vectors. This matrix will have the n true location vectors as its argument, apart from any parameters in the covariance function. Implementing a Markov chain Monte Carlo algorithm to extract posterior samples to learn about the true positional errors embedded in the likelihood in this fashion will be hopeless—the identifiability will be very weak and the computational burden will be very high. Attempting to fit the marginal model resulting from marginalisation over the true location vectors is even less promising—no explicit integration is possible. The measurement error model specification would make matters even worse. Now we would require a prior over the n true locations. This prior can not involve any map information—the best we could envision would be a uniform prior over the entire map, further weakening the identifiability of the true locations. In summary, the model of the previous paragraph appears to offer the only viable option and so we confine ourselves to it in the sequel.

A related point arises here. With multiple maps, we propose registering each one separately with regard to the GPS data, i.e., a different regression model for each map. Then, we suggest Bayesian model averaging to perform the predictive inference for true location, hence for positional error associated with any of the maps. That is, each map is equivalent to a regression model so model averaging is map averaging. However, with say L maps, for map $i, i = 1, 2, \ldots, L$, the collection of feature locations associated with the GPS locations, say x_i, can be viewed as one of L covariates to explain the GPS locations. Why not use all of the maps to explain the GPS locations and thus to do predictive inference on true locations, rather than using model averaging? Of course this is computationally feasible but there will be high multicollinearity in the x_i—after all, they are supposed to be the same set of locations—resulting in severe over-fitting and very inflated predictive variances. Essentially, a single covariate model is likely to be among the best in avoiding over-fitting. Indeed, the proposed model averaging averages over such models, yielding prediction with smaller uncertainty than that associated with any individual model (map) and so is expected to be more attractive than a "multiple regression" approach.

We implement our approach within a Bayesian framework using hierarchical modelling, enabling a fully model-based examination of positional error. In terms of measurement error settings, hierarchical modelling is discussed in, e.g., Gilks, Richardson & Spiegelhalter (1996)

and in Gelfand & Mallick (1996). Our work may also play the role of a precursor to fully model-based development of error propagation studies (Heuvelink, Burrough & Stein 1989; Heuvelink 1999). Typically, these studies show how a priori positional uncertainty, often determined from a map's metadata, is propagated through various GIS operations on digital maps (Hunter & Goodchild 1996; Stanislawski, Dewitt & Shrestha 1996). Like these existing methods, our method uses map positional error metadata and documented GPS performance to help specify prior uncertainty with regard to both spatial and non-spatial error. We combine this prior information with map and control point data via a Bayesian hierarchical model to obtain full posterior inference regarding the true location associated with any map point, e.g., a point estimate and a two-dimensional credible set. For a road, we can simulate posterior realisations of the true road, for a city block, posterior realisations of the true block, again to achieve point estimation and to assess variability.

We develop the model for a single map in Section 2, detailing prediction in Section 3. We extend the model to the case of multiple maps using Bayesian model averaging in Section 4. Section 5 turns to an application working with three neighbourhood maps in Durham, North Carolina, along with GPS location data. We illustrate inference for point, road and city block features. We conclude with a brief discussion in Section 6, noting problems to be considered for future work in this area.

2. A MAP POSITIONAL ERROR MODEL

Over the following three subsections we assemble the proposed positional error model.

2.1. Notation.

Consider L maps or data layers of some region, $\mathcal{X}$. Our interest centres on characterising the positional error of features that are represented by coordinates on one or more maps. We discuss positional errors of flat-map coordinates, so in this article, $\mathcal{X}$ will be taken to be a subset of some 2-dimensional coordinate system.

For example, a road intersection or a marker such as a utility pole is considered a point feature in that its location is represented on a map as a single coordinate pair. One dimensional features, e.g., roads and rivers, as well as two dimensional areal features, e.g., a property boundary, are viewed the way they are stored in GIS software, through polygonal curves defined by the points that determine the segments of the curves.

An important issue for us is notation. Following the Introduction, we have to denote the following types of objects: map locations, control (GPS) locations, and true locations. GPS locations and true locations are one-to-one but are not connected to any particular map. However, for a given true location, hence GPS location, we potentially have L different map locations. We address this situation by considering map-specific transformations from map locations to associated GPS locations, and similarly from map to true locations. We use x to denote an arbitrary point on an arbitrary map, occasionally using $x = (x_1, x_2)^T$ to make explicit reference to the east-west and north-south coordinates. In customary regression notation, we denote the transformations for map i to GPS locations and to true locations by $y_{\mathrm{gps},i}(x)$ and $y_{\mathrm{true},i}(x)$, respectively. If we assume that x, $y_{\mathrm{gps},i}(x)$ and $y_{\mathrm{true},i}(x)$ are in a common reference system (e.g., UTM NAD 83), then the true positional error vector associated with location x on map i is defined as $y_{\mathrm{true},i}(x) - x$. Again, following the discussion of the Introduction, in asking about positional error, x will be provided, so our focus is on modelling $y_{\mathrm{true},i}(x)$. We note that our use of the terms "true" or "truth" will refer to projected coordinates and not to coordinates for an unprojected geodetic location, which spatial scientists more commonly refer to as "truth."

Though a map displays an uncountable number of locations, it is stored as a finite set of locations or features. Hence, we assume map $i, i = 1, 2, \ldots, L$, to have N_i points, $x_{ij}, j = 1, \ldots, N_i$. In practise, the corresponding observed number of control points is usually much smaller than N_i. So, we denote the actual number of pairs of observed map locations and GPS

locations for map i by n_i and write these pairs as $(x_{ij}, y_{\mathrm{gps},ij})$, i.e., $y_{\mathrm{gps},ij}$ is an abbreviation for $y_{\mathrm{gps},i}(x_{ij})$. These pairs are used to fit the registration model for map i, but of course we seek to learn about positional error for any of the x_{ij}. We remark that, for maps i and i', if x_{ij} and $x_{i'j'}$ denote the same feature, then $y_{\mathrm{gps},ij} = y_{\mathrm{gps},i'j'}$.

2.2. Prior information.

As noted in Section 1, the coordinates of the control points are not viewed as the true locations but still have measurement error. Under the Berkson model, we write this as $y_{\mathrm{true}} = y_{\mathrm{gps}} + \eta$ where the η are independent and identically distributed bivariate normal with mean 0 and covariance matrix $\sigma_{\mathrm{g}}^2 I_2$, where I_2 is the 2×2 identity matrix, and σ_{g}^2 is essentially known. Such a specification is suggested by summaries of GPS data as discussed in the Trimble GPS documentation (Trimble Navigation 1997a, b, 2003). While it is likely that in the covariance matrix for η, the NS $\sigma^2 \neq$ EW σ^2 and that the off-diagonal entry is non-negligible (Soler & Marshall 2002), in the absence of measured truth and given that we anticipate measurement error to be a small component of map positional error, we retain this simple specification.

The Berkson specification is a conditional model, i.e., conditional on y_{gps} and, in the sequel, we interpret σ_{g}^2 in this fashion. Also, it is evident that this specification is not associated with any map. However, in terms of locations on map i, we may write the relationship as

$$(y_{\mathrm{true},i}(x) - y_{\mathrm{gps},i}(x)) \,|\, y_{\mathrm{gps},i}(x) \sim \mathsf{N}(0, \sigma_{\mathrm{g}}^2 I_2).$$

But then, under a registration model for $y_{\mathrm{gps},i}(x)$ given x such as we develop in Section 2.3, marginalisation over $y_{\mathrm{gps},i}(x)$ yields a model for $(y_{\mathrm{true},i}(x) - x) \,|\, x$, the desired positional error model for map i.

Maps typically come with positional uncertainty metadata, e.g., an associated RMSE (root mean squared error) accuracy which provides prior information on the variance for this marginal model. As an example of such information, consider the Digital Line Graphs (DLGs) produced by the US Geological Survey (USGS). Locations on large-scale DLGs "... shall be less than or equal to 0.003 inches standard error in both the x and y component directions, relative to the source that was digitized" (Digital Line Graphs Standards, http://rockyweb.cr.usgs.gov/nmpstds/dlgstds.html (04/28/04)), the "standard error" simply being the square root of the mean squared differences between coordinates on the DLG and the corresponding source map coordinates.

The source of most large-scale DLGs consists of USGS 7.5 minute topographic quadrangle maps having a scale of 1:24000 and are stated to comply with National Map Accuracy Standards so that "... not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50 inch" (National Map Accuracy Standards, http://rockyweb.cr.usgs.gov/nmpstds/nmas.html (04/28/04)). This translates to 40ft at the scale of 1:24000. If we assume a bivariate Gaussian distribution for the total positional error vector having uncorrelated components with common scale, this corresponds to the 90% error circle of $\mathsf{N}(0, (18.64ft)^2 I_2)$. At the same map scale, the DLG production process nominally introduces a standard error less than 6ft. Thus, if we are willing to assume that error components are uncorrelated normal variates with common variance, and if we consider the DLG production process to be independent of the production of quadrangle maps, then we might conclude that a large-scale DLG has a variance less than about $20^2\mathrm{ft}^2$ $(18.64^2 + 6^2)$ or about $6^2\mathrm{m}^2$. In the sequel, we denote this a priori total positional error vector variance component for map i as $\sigma_{t,i}^2$. Of course, the model we work with for $(y_{\mathrm{true},i}(x) - x) \,|\, x$ incorporates a general covariance matrix for total positional error.

2.3. Model development.

With fitting data for map i denoted by $(y_{\mathrm{gps},ij}, x_{ij})$, $j = 1, \ldots, n_i$, a typical procedure for adjusting map points given a set of control points would first fit a global model for the $y_{\mathrm{gps},ij}$ to

the x_{ij}, so-called *registration*, yielding fitted values, say, $\hat{y}_{ij} \equiv \hat{y}_i(x_{ij}), j = 1, 2, \ldots, n_i$, and, in fact, for any feature x on map i the predictor $\hat{y}_i(x)$. The six-parameter affine transformation below is commonly used in this context—though our modelling formulation can be used with other choices of transformation; and, in anticipation of the notation used in our fully developed model below, we denote it as $\mu_i(x_{ij}) \equiv (\mu_{i1}(x_{ij}), \mu_{i2}(x_{ij}))^T$, where

$$\begin{aligned} \mu_{i1}(x_{ij}) &= \beta_{i10} + \beta_{i11}x_{ij1} + \beta_{i12}x_{ij2}, \\ \mu_{i2}(x_{ij}) &= \beta_{i20} + \beta_{i21}x_{ij1} + \beta_{i22}x_{ij2}; \end{aligned} \tag{1}$$

see, *e.g.*, Dowman (1999). In a typical GIS implementation, the β parameters are fitted via least squares, and the goodness of the transformation is typically reported as the root mean squared error,

$$\left(\frac{1}{n_i} \sum_{j=1}^{n_i} (y_{\mathrm{gps},ij} - \hat{y}_{ij})^T (y_{\mathrm{gps},ij} - \hat{y}_{ij}) \right)^{\frac{1}{2}}.$$

A transformation like (1) may be considered a large-scale adjustment to make a map to match the control points in some overall average sense—rotation, scaling, shifting.

The next step in the adjustment procedure is referred to as "rubber-sheeting", i.e., local adjustment (White & Griffin 1985), which is apart from registration, and results in revising $\hat{y}_i(x)$ to, say, $\hat{\hat{y}}_i(x)$. This informal procedure is not model-based; the function $\hat{\hat{y}}_i(\,\cdot\,)$ is not explicitly defined. However, this procedure does reflect the expectation that, after large-scale adjustment, small scale misalignments will still exist. Since this informal procedure is inherently spatial, in moving to a formal model to capture this residual misalignment, it is appropriate to introduce spatial structure. Additionally, we may choose to constrain some of the transformed x_{ij} to match exactly the corresponding $y_{\mathrm{gps},ij}$, with remaining points on map i being pulled along in some local way. This is inherent in rubber-sheeting procedures (White & Griffin 1985) in GIS whereby transformations are applied piecewise to local regions defined by some tessellation of the map where the constraints occur at the vertices of the tessellation tiles.

Hence for each x_{ij} we have

$$\begin{aligned} y_{\mathrm{gps},ij} &= \hat{y}_{ij} + (y_{\mathrm{gps},ij} - \hat{y}_{ij}) \\ &= \hat{y}_{ij} + (\hat{\hat{y}}_{ij} - \hat{y}_{ij}) + (y_{\mathrm{gps},ij} - \hat{\hat{y}}_{ij}), \end{aligned}$$

where $\hat{\hat{y}}_{ij} \equiv \hat{\hat{y}}_i(x_{ij})$. So in explaining $y_{\mathrm{gps},ij}$, we have a global estimate, $\hat{y}_{ij}$, a rubber-sheeting adjustment, $(\hat{\hat{y}}_{ij} - \hat{y}_{ij})$, and a residual error, $(y_{\mathrm{gps},ij} - \hat{\hat{y}}_{ij})$.

As an explanatory model for $y_{\mathrm{gps},i}(x)$, this motivates, the GPS location associated with map location x on map i,

$$y_{\mathrm{gps},i}(x) = \mu_i(x) + v_i(x) + \varepsilon_i(x), \tag{2}$$

where $\mu_i(x)$ is a parametric global surface for map i, $v_i(x)$ is a bivariate spatial process—details below—to provide a model-based adjustment to the global estimate rather than an ad hoc one, and $\varepsilon_i(x)$ is a non-spatial or pure noise process. As alluded to above, we use (1) for $\mu_i(x)$.

The specification (2) does not force revised map locations to exactly match the associated observed GPS locations at, say, a set of vertices as in a typical rubber-sheeting algorithm. However, if we set $\varepsilon_i(x) = 0$, and take $v_i(x)$ as a mean-square continuous bivariate spatial process, then the model behaves much like rubber-sheeting. That is, if we observe $(y_{\mathrm{gps},i}(x_{ij}), x_{ij})$, $j = 1, \ldots, n_i$, in the absence of pure error, the interpolated position for each x_{ij} is exactly $y_{\mathrm{gps},i}(x_{ij})$; the model "honor[s] the data" just as in (co-)kriging without measurement error (Cressie 1993, § 3.2.1). And prediction of $y_{\mathrm{gps},i}(x_0)$ for some map point x_0 is most influenced by the control points $y_{\mathrm{gps},i}(x_{ij})$ corresponding to the x_{ij} closest to x_0. Thus, notions of "pinning" and local adjustment are preserved as in rubber-sheeting. Adding the noise term $\varepsilon_i(x)$ enables residual explanation that need not be entirely spatial. This is arguably more flexible and more realistic. In practise, approximate pinning will occur; see Figure 2 below.

We use the flexible linear model of coregionalization (Wackernagel 2003) to provide an isotropic specification for the spatial dependence in the $v_i(x)$, hence in the $y_{\mathrm{gps},i}(x)$. In particular, we employ a version of the linear model of coregionalization developed in Banerjee, Carlin & Gelfand (2004, § 7.2). Let $w_i(x) = (w_{i1}(x), w_{i2}(x))^T$, where $w_{i1}(x)$ and $w_{i2}(x)$ are uncorrelated spatial processes with mean zero, unit variance and spatial correlation functions ρ_{i1} and ρ_{i2}. For convenience, we assume ρ_{ik} is an exponential correlation function with scalar range parameter, ϕ_{ik}. Coregionalization creates a bivariate spatial process by linear transformation of two independent univariate processes. More precisely, we model $v_i(x)$ via the linear model of coregionalization as

$$v_i(x) = A_i w_i(x)$$

where A_i is the map-specific 2×2 coregionalization matrix providing the unknown linear transformation and without loss of generality, can be taken to be lower triangular. Thus, coordinatewise, we have

$$\begin{aligned} v_{i1}(x) &= a_{i11} w_{i1}(x), \\ v_{i2}(x) &= a_{i21} w_{i1}(x) + a_{i22} w_{i2}(x). \end{aligned}$$

We take the coordinates of the map-specific pure error $\varepsilon_i(x)$ to be mutually and internally uncorrelated error processes with scale parameters $\sigma_{\varepsilon_i k}$, $k = 1, 2$.

In this model we have

$$\begin{aligned} v_i(x) &\sim \mathsf{N}(0, T_i), \quad \text{and} \\ \varepsilon_i(x) &\sim \mathsf{N}(0, G_i), \end{aligned}$$

where $T_i = A_i A_i^T$, and G_i is the diagonal matrix with components $\sigma^2_{\varepsilon_i k}$, $k = 1, 2$. So, the variance associated with a GPS location as explained under the model for map i is

$$\Sigma_{y_{\mathrm{gps},i}(x)} = T_i + G_i.$$

But then since $\Sigma_{y_{\mathrm{true},i}(x)} \,|\, y_{\mathrm{gps},i}(x) = \sigma^2_{\mathrm{g}} I_2$, we have

$$\Sigma_{y_{\mathrm{true},i}(x)} = \sigma^2_{\mathrm{g}} I_2 + T_i + G_i,$$

the marginal positional error covariance alluded to earlier. Evidently, in offering (2), along with a GPS error model, as a fully specified stochastic analogue to the usual map positional error assessment, our modelling has yielded a more complex positional error structure than $\sigma^2_{\mathrm{t},i} I_2$, where $\sigma^2_{\mathrm{t},i}$ corresponds to the total positional uncertainty derived from the metadata for map i (See Section 2.2). However, the latter can be used to help with the prior specifications by viewing

$$\begin{aligned} t_{i11} + \sigma^2_{\varepsilon_i 1} &\approx \sigma^2_{\mathrm{t},i} - \sigma^2_{\mathrm{g}}, \\ t_{i22} + \sigma^2_{\varepsilon_i 2} &\approx \sigma^2_{\mathrm{t},i} - \sigma^2_{\mathrm{g}}, \end{aligned} \tag{3}$$

where t_{i11} and t_{i22} are the diagonal elements of T_i. Thus a priori we centre each of t_{i11}, t_{i22}, $\sigma^2_{\varepsilon_i 1}$ and $\sigma^2_{\varepsilon_i 2}$, around $\frac{1}{2}(\sigma^2_{\mathrm{t},i} - \sigma^2_{\mathrm{g}})$.

Such a prior suggests a neutral opinion regarding the relative size of the spatial and pure error contributions. See Section 5.1 for detailed prior specification. Following the discussion above, we might specify the prior to encourage a relatively larger variance component for the t since rubber-sheeting (as an ad hoc spatial operation) attempts to remove the pure error and pins $\hat{\hat{y}}_{ij}$ to $y_{\mathrm{gps},ij}$ at the control points. However, as we argued above, this does not suggest removing $\varepsilon_i(x)$. Ad hoc rubber-sheeting will not be perfect for all x, so by analogy, why should we insist that the spatial process correction be perfect? The prior also suggests no opinion regarding differential uncertainty according to direction.

3. PREDICTION OF TRUE LOCATION

Again, our primary goal is the prediction of the true location associated with location, say x_0 on map i. As a result, we seek the predictive distribution for $y_{\mathrm{true},i}(x_0)$; we would not be interested in $y_{\mathrm{gps},i}(x_0)$. In particular, our point estimate is

$$\begin{aligned} \mathrm{E}\left(y_{\mathrm{true},i}(x_0) \mid D_i\right) &= \mathrm{E}\left(y_{\mathrm{gps},i}(x_0) \mid D_i\right) \\ &= \mathrm{E}\,\mathrm{E}\left(y_{\mathrm{gps},i}(x_0) \mid \mu_i(x_0), v_i(x_0), D_i\right) \\ &= \mathrm{E}\left(\mu_i(x_0) + v_i(x_0) \mid D_i\right), \end{aligned} \tag{4}$$

where D_i is the data, $\{(y_{\mathrm{gps},ij}, x_{ij})\, j = 1, \ldots, n_i\}$. However, uncertainty is decreasing as we progress from $[y_{\mathrm{true},i}(x_0) \mid D_i]$ to $[y_{\mathrm{gps},i}(x_0) \mid D_i]$ to $[\mu_i(x_0) + v_i(x_0) \mid D_i]$, using $[\,\cdot\,]$ to denote a density or mass function.

To quantify the uncertainty associated with each of these distributions, we would obtain posterior samples using the usual composition with posterior draws of the β_i, the T_i and the G_i. In particular, we want to sample from

$$\left[y_{\mathrm{true},i}(x_0) \mid \sigma^2_{\mathrm{g}}, D_i\right] = \int \left[y_{\mathrm{true},i}(x_0) \mid y_{\mathrm{gps},i}(x_0), \sigma^2_{\mathrm{g}}\right] \cdot \left[y_{\mathrm{gps},i}(x_0) \mid D_i\right] \mathrm{d}y_{\mathrm{gps},i}, \tag{5}$$

but

$$\left[y_{\mathrm{gps},i}(x_0) \mid D_i\right] = \int \left[y_{\mathrm{gps},i}(x_0) \mid \mu_i(x_0), \Sigma_{y_{\mathrm{gps},i}(x_0)}\right] \cdot \left[\{\beta_{ikl}\}, T_i, G_i \mid D_i\right] \mathrm{d}\theta_i,$$

where we use θ_i generically to denote the vector of posterior parameter values over which the integration occurs.

For a linear feature, e.g., the line on map i between x_0 and x_0^*, we will obtain a posterior sample of lines by connecting the sampled pairs $y_{\mathrm{true},i}(x_0)$ and $y_{\mathrm{true},i}(x_0^*)$. For a city block, similarly, we will obtain a posterior sample of rhombi. Note that sampling the posterior predictive distribution $\left[y_{\mathrm{true},i}(x_0) \mid \sigma^2_{\mathrm{g}}, D_i\right]$ can be done after we have fitted the model to explain $y_{\mathrm{gps},i}(x)$ for map i, i.e., after we have collected the posterior samples of the model parameters. It is also clear that each map model is fitted separately. There are no common parameters across models, and there is no natural additional level of hierarchical specification to link the models.

With n_i fitting points x_{ij} for map i, $i = 1, \ldots, L$, and with the assumption of Gaussian distributions for all processes, we write $\tilde{v}_i = (v_i^T(x_{i1}), \ldots, v_i^T(x_{in_i}))^T$ so that

$$\tilde{v}_i \sim \mathsf{N}\left(0, \sum_{k=1}^{2} R_{ik} \otimes T_{ik}\right),$$

where R_{ik} is an $n_i \times n_i$ matrix with entries $\rho_{ik}(x_{ij} - x_{ij'})$, $j, j' = 1, \ldots, n_i$, and $T_{ik} = a_{ik} a_{ik}^T$, where a_{ik} is the k^{th} column of A_i and $\otimes$ denotes the Kronecker product; see Banerjee, Carlin & Gelfand (2004, § 7.2). With $\tilde{y}_{\mathrm{gps},i}$ and $\tilde{\mu}_i$ defined analogously to $\tilde{v}_i$, the conditional distribution of $\tilde{y}_{\mathrm{gps},i}$ given $\tilde{\mu}_i$ and $\tilde{v}_i$ is

$$\tilde{y}_{\mathrm{gps},i} \mid \tilde{\mu}_i, \tilde{v}_i \sim \mathsf{N}(\tilde{\mu}_i + \tilde{v}_i, I_{n_i \times n_i} \otimes G_i).$$

Assuming prior independence of parameters, we have the probability model for map i,

$$\left[\tilde{y}_{\mathrm{gps},i}, \mid \tilde{v}_i, \{x_{ij}\}, \{\beta_{ikl}\}, G_i\right] \cdot \left[\tilde{v}_i \mid \{\phi_{ik}\}, T_i\right] \cdot [T_i] \cdot [G_i] \prod_{k=1}^{2} [\phi_{ik}] \cdot [\tilde{\beta}_{ik}],$$

where, again, $T_i = A_i A_i^T = \sum_{k=1}^{2} a_{ik} a_{ik}^T = \sum_{k=1}^{2} T_{ik}$, and $\tilde{\beta}_{ik} \equiv (\beta_{ik0}, \beta_{ik1}, \beta_{ik2})^T$. Integrating over $\tilde{v}_i$ gives,

$$\tilde{y}_{\mathrm{gps},i} \sim \mathsf{N}\left(\tilde{\mu}_i, \sum_{k=1}^{2} R_{ik} \otimes T_{ik} + I_{N_i \times N_i} \otimes G_i\right),$$

and the marginal model of the form

$$\big[\tilde{y}_{\mathrm{gps},i}|\,\{x_{ij}\},\{\beta_{ikl}\},\{\phi_{ik}\},T_i,G_i\big]\cdot[T_i]\cdot[G_i]\prod_{k=1}^{2}[\phi_{ik}]\cdot[\tilde{\beta}_{ik}]. \tag{6}$$

We assign weak independent scaled inverse-chi-square distributions for the pure error variances $\sigma^2_{\varepsilon_i k}$ in G_i, centred around the prior information—see (3) and surrounding text—with infinite variance. Vague independent normal distributions are assigned to the parameters β_{ikl} of the affine transformation (1). Independent gammas are assigned to the range parameters ϕ_{ik}, and we use Metropolis–Hastings steps for sampling. An informative inverse-Wishart prior is used for the coregionalization matrix T_i. We use Metropolis–Hastings steps for each of the components a_{i11}, a_{i22}, and a_{i21} of the matrix A_i; we use lognormal proposals for each of first two and a normal proposal for the latter. These require a factor of $|J|$, where $J = 4a^2_{i11}a_{i22}$ is the Jacobian of the transformation from A_i to T_i. The random effects $\tilde{v}_i$ are sampled from the conjugate normal that results. We specify the particular priors in Section 5.

Alternatively, we could have used (6) to avoid sampling the random effects and then used Metropolis–Hastings steps for the $\sigma^2_{\varepsilon_i k}$ with a similar sampling scheme as above for the remaining quantities. Regardless, we can sample the predictive distribution for $y_{\mathrm{gps},i}(x_0)$ after the Markov chain Monte Carlo routine in a one-for-one fashion. Then, sampling from $\big[y_{\mathrm{true},i}(x_0)\,|\,y_{\mathrm{gps},i}(x_0),\sigma^2_{\mathrm{g}}\big]$ yields posterior draws for $y_{\mathrm{true},i}(x_0)$ and subtracting x_0 for the positional error at x_0.

4. BAYESIAN MODEL AVERAGING FOR MULTIPLE MAPS

At this point, for each map i, $i = 1,\ldots,L$, we have a fitted model for the collection of n_i observed control points $\tilde{y}_{\mathrm{gps},i}$ with the corresponding collection of map points, $\tilde{x}_i$. Beyond handling positional error for a given map, we may be interested in improving the prediction of true locations by using all of the maps. Indeed, true location is the same regardless of map model suggesting the use of Bayesian model averaging (Raftery, Madigan & Hoeting 1997). In order to glean any information about a particular true location from map i, however, we must know the map feature location on map i associated with this true location. That is, customary Bayesian model averaging assumes that we work with the same data for each model. In the multiple map case, feature locations change across maps, and we may have maps that do not share features with other maps.

Suppose that $y_{\mathrm{true},0}$ is associated with a feature location on all maps, say x_{i0} for map i; if it is associated with a location only on a subset of the L maps, we will use that subset to do the model averaging. Let M_i denote the map-model for map i as developed in Section 2, and let $[M_i]$ be the a priori probability mass for map-model M_i. We can use a uniform prior over the maps or perhaps one that reflects our prior confidence in the maps, say using the map root mean squared error in some fashion. The predictive distribution of $y_{\mathrm{true},0}$ follows in the standard way from an average of the predictive distributions for each model weighted by the posterior model probability,

$$\big[y_{\mathrm{true},0}\,|\,\sigma^2_{\mathrm{g}},D\big]=\sum_{i=1}^{L}\big[y_{\mathrm{true},i}(x_{i0})\,|\,\sigma^2_{\mathrm{g}},D,M_i\big]\cdot\big[M_i\,|\,D\big],$$

where, though we might write D_i to denote the data associated with fitting model i, it is notationally convenient to simply let D denote $\tilde{y}_{gps}$ since we are thinking of the $\tilde{x}_i$ as fixed. The first factor in the summand is given by (5), and the posterior probability for model M_i is

$$[M_i\,|\,D]=\frac{[D\,|\,M_i]\cdot[M_i]}{\sum_{l=1}^{L}[D\,|\,M_l]\cdot[M_l]};$$

and

$$[D\,|\,M_i]=\int[D\,|\,\theta_i,\,M_i]\cdot[\theta_i\,|\,M_i]\,\mathrm{d}\theta_i$$

is the integrated likelihood for model M_i, θ_i is the collection of all parameters for model M_i with $[\theta_i \,|\, M_i]$ being the joint prior, and $[D \,|\, \theta_i,\, M_i]$ is the likelihood for map-model i, i.e., $[D \,|\, \theta_i,\, M_i] = \left[\tilde{y}_{\rm gps}|\tilde{x}_i, \theta_i, M_i\right]$.

This model averaging may be interpreted as averaging over different regression models. Each regression model is associated with a map. Each regression model uses a different covariate. The covariate for a particular regression model is the set of map locations used for the associated map. The approach of Chib (1995), more precisely Chib & Jeliazkov (2001) can be used to compute marginal likelihoods. Simultaneous prediction at a collection of locations follows similarly, only requiring sampling from conditional multivariate Gaussian distributions.

5. APPLICATION

We illustrate our approach using three maps of a residential neighbourhood in Durham, North Carolina (Figure 1).

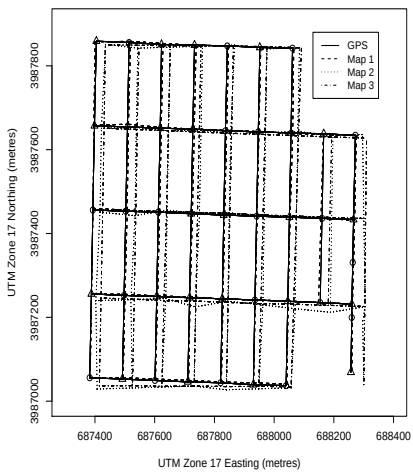

FIGURE 1: Abridged map of a neighbourhood in west Durham, NC, showing, under common reference, the map derived from GPS coordinates and three maps from existing spatial databases. 20 circles indicate (GPS) control points used for model fitting, and 24 triangles indicate remaining GPS points.

One ($i = 1$) is a large-scale (1:24000) USGS Digital Line Graph (DLG) (http://edcftp.cr.usgs.gov/pub/data/DLG/LARGE_SCALE/N/northwest_durham_NC/transportation/867112.RD.sdts.tar.gz (01/20/04)), another ($i = 2$) from a Census 2000 TIGER/Line database (http://esri.com/data/download/census2000_tigerline/index.html, North Carolina, Durham County, Line Features-Roads (12/17/03)), and the third ($i = 3$) from StreetMap USA (Environmental Systems Research Institute 2003), an enhanced version of the Census 2000 TIGER/Line database. Each map shares the same 44 street intersections for which we also obtained control points using the mean of approximately 200 differentially corrected GPS coordinates for each intersection. All maps and control points were transformed to a common reference system (NAD 83, UTM metres) before analysis. Under the common reference system, Figure 1 shows the three maps, the GPS-based map, and the 44 GPS points. To work with smaller eastings and northings, all coordinates, including GPS, were then centred by the overall mean coordinates of all three maps, not including GPS. We used a common subset of $n_i = n = 20$ control points and corresponding map points to fit the three models and for each we generated posterior predictions of true values using (5) at the remaining 24 locations.

5.1. Priors and starting values.

Prior map information suggests that the DLG ($i = 1$) has the highest positional accuracy, while the TIGER/Line file ($i = 2$) and the StreetMap USA file ($i = 3$) may be expected to have similar accuracy. We use the National Map Accuracy Standards as metadata for source map accuracy (http://rockyweb.cr.usgs.gov/nmpstds/nmas.html (04/28/04)) and use the DLG Standards as metadata for production accuracy of DLGs from source maps (http://rockyweb.cr.usgs.gov/nmpstds/dlgstds.html (04/28/04)). Assuming that source map errors and production errors are independent allows us to express a priori the source map and production accuracy metadata as the variance of a circular bivariate Gaussian distribution of positional error and suggests that the DLG has prior accuracy corresponding to a variance of $\sigma^2_{\mathrm{t},1} \approx 6^2\mathrm{m}^2$; see Section 2.2.

The prior information is less clear for the TIGER/Line and the StreetMap USA databases. The source maps for these two particular databases are uncertain, but we rely on the fact that many of the sources for these databases have a scale of 1:100000; we assume this scale for the source of these files. This implies a prior variance of about $24^2\mathrm{m}^2$. Also, although we do not know the production standards as in the case of the DLG, we still assume "0.003 inches standard error" due to the production process from the source; this implies a variance of about $8^2\mathrm{m}^2$ at 1:100000. Together these suggest that a priori $\sigma^2_{\mathrm{t},2} = \sigma^2_{\mathrm{t},3} \approx 25^2\mathrm{m}^2$. The processing software for GPS positions indicates a positional accuracy corresponding to a variance of about $\sigma^2_{\mathrm{g}} = 0.7^2\mathrm{m}^2$, which we assume is known (Trimble Navigation 1997a, b, 2003); some exploratory analysis of the GPS positions suggests that this is a reasonable value.

Thus, according to the discussion near (3), we centre the priors for each $\sigma^2_{\varepsilon_i k}$ at means of $\frac{1}{2}6^2$, $\frac{1}{2}24^2$ and $\frac{1}{2}24^2$ for $i = 1, 2, 3$, respectively, $k = 1, 2$. We take the degrees of freedom to be 4, the same as that given to the inverse-Wishart prior for the T_i coregionalization matrices; see below. As a result, the priors corresponding to the above means and degrees of freedom are $\mathrm{Inv}-\chi^2(4, 9)$, $\mathrm{Inv}-\chi^2(4, 144)$, and $\mathrm{Inv}-\chi^2(4, 144)$, where $\mathrm{Inv}-\chi^2(\,\cdot\,,\cdot\,)$ denotes the scaled inverse-chi-squared distribution. These priors are weak in that they have infinite variance. The parameterisations of all distributions follow Gelman, Carlin, Stern & Rubin (1995).

We use inverse-Wishart priors for the coregionalization matrices $T_i \sim IW_{\nu_i}(S_i^{-1})$, $i = 1, 2, 3$, where $\nu_i = 4$, degrees of freedom, and, again, according to the discussion near (3), $S_1 = \frac{1}{2}6^2 I_2$ and $S_2 = S_3 = \frac{1}{2}24^2 I_2$ where S_i is the scale matrix of the inverse-Wishart; $\nu_i > 3$ is required for a proper distribution in two dimensions. We adopt a vague normal distribution for the parameters of the affine transformation with mean $\mathrm{E}\,(\beta_{i10}, \beta_{i11}, \beta_{i12}, \beta_{i20}, \beta_{i21}, \beta_{i22})^T = (0, 1, 0, 0, 0, 1)^T$ and variance $100^2 I_2 \otimes (X_i^T X_i)^{-1}$, where X_i denotes the $n_i \times 3$ regression design matrix formed by augmenting a column of ones with the two columns of the centred map coordinates for map i. Finally we take independent gamma distributions $\Gamma(1, 0.00575)$ for the range parameters ϕ_{ik}. This corresponds to a mode of 0.00575 at 0m, tapering to 0.000575 at 400m, roughly half the diameter of the study area.

Starting values for the parameters and spatial random effects were determined by using the GPS control points and map points as data in maximum likelihood estimation applied separately to each coordinate. With regard to sensitivity, we tried other starting values with no practical effect on the final results.

5.2. Results.

To assess convergence of three chains for each map model, we used the potential scale reduction factor (Gelman & Rubin 1992) and its multivariate version (Brooks & Gelman 1997), as implemented in the CODA add-on package of the R Package for Statistical Computing (R Development Core Team 2004). Convergence was achieved if each of the univariate factors and the multivariate version were less than 1.1. This occurred between 5000 and 10000 iterations. We continued sampling one chain for another 50000 iterations and present results based on every 50th draw beyond 10000 for a total of 1000 samples from the posterior. Posterior summaries for each map are given in Tables 1–3.

Although we are primarily interested in positional error, examination of the tables reveals several interesting points. First, as anticipated, map 1 is superior with regard to the extent of uncertainty. Maps 2 and 3 are similar in their performance. Indeed their global (affine) transformations are essentially the same and both are quite different from that of map 1. For all three maps, spatial variation is roughly of the same magnitude as pure error variation. Evidently, a pure error component is needed; a purely spatial model would not be adequate; i.e., $\varepsilon_i(x)$ is not zero. Association between the north-south spatial correction and the east-west spatial correction does not emerge as significant for any of the maps as indicated by zero being contained within 95% credible intervals of each of the marginal posterior distributions of the t_{i21}, $i = 1, 2, 3$. For each map we can compare the prior uncertainty ($\sigma^2_{\mathrm{t},i} - \sigma^2_{\mathrm{g}}$) with the estimated uncertainty, the posterior mean of $t_{ijj} + \sigma^2_{\varepsilon_i,j}$; see (3). The magnitude of the posterior mean was about 35% of the corresponding prior quantity for map 1 and about 20% for maps 2 and 3, each prior value exceeding its corresponding posterior 95% credible interval; evidently, some of the posterior variability is explained by the large-scale transformation. Figure 2 attempts to demonstrate the approximate pinning that results from the map models. In particular it shows again the GPS coordinates and the estimated coordinates under map 1, the latter using expression (4).

TABLE 1: DLG map model ($i = 1$) posterior summary.

Name	2.5%	50%	Mean	97.5%	IQR
β_{110}	-5.77464	-2.41488	-2.47180	0.39073	1.61296
β_{111}	0.99473	1.00238	1.00223	1.00938	0.00474
β_{112}	-0.01048	-0.00317	-0.00324	0.00385	0.00474
β_{120}	-4.48818	-1.65955	-1.67575	0.99796	1.57486
β_{121}	-0.00566	0.00139	0.00115	0.00739	0.00402
β_{122}	0.99546	1.00190	1.00196	1.00836	0.00408
t_{111}	2.24364	5.84744	6.76246	16.47642	3.98051
t_{122}	2.01010	4.69761	5.38192	12.31547	3.09828
t_{121}	-3.89903	0.31556	0.26765	4.35008	2.20939
ϕ_{111}	6.43267	206.74364	265.39230	847.37141	281.39840
ϕ_{112}	8.39029	295.97271	339.80958	980.00578	340.93933
$\sigma^2_{\varepsilon_1 1}$	3.92300	7.71274	8.37712	16.98480	3.82208
$\sigma^2_{\varepsilon_1 2}$	2.74388	5.26588	5.74685	11.29316	2.61991

Turning to Bayesian model averaging, the log marginal likelihood values for the three maps are -125.303 ($i = 1$), -165.220 ($i = 2$), and -165.260 ($i = 3$). At these magnitudes, the posterior probability on model 1, $[M_1 \mid D]$, becomes essentially 1 under any reasonable prior masses $[M_i]$, $i = 1, 2, 3$. Although this result is extreme and renders model averaging uninteresting, it agrees with our prior information about map accuracy and with the ensuing data analysis regarding map positional accuracy in Tables 1–3. Furthermore, the log integrated likelihood values may be viewed as comparative summary measures of positional accuracy among maps.

With regard to posterior prediction for the true positions $y_{\mathrm{true}}(x_0)$ corresponding to the 24 points not used in the modelling procedure, for each map 1000 posterior draws were taken and are summarised in the left column of Figure 3, which shows the 95% credible ellipses for true position corresponding to a bivariate normal approximation with mean and covariance computed from the draws. The GPS locations are marked with a $+$, map points with a $\times$. The greater uncertainty attached to maps 2 and 3 is again revealed. Moreover, the ellipses do not include several of their respective locations on maps 2 and 3; we note that this not an indication of poor

model performance but rather of poor map accuracy relative to our ability to predict true location with our models. As an indication of our models' predictive performances, we also computed the percentage of the 24 GPS points that fall within their respective GPS credible ellipses (not shown); all 24 GPS points on each of maps 2 and 3 fall within their respective approximate 95% credible ellipses, and all but 1 on map 1—upper left +—fall in their respective ellipses.

TABLE 2: Census 2000 TIGER/Line map model ($i = 2$) posterior summary.

Name	2.5%	50%	Mean	97.5%	IQR
β_{210}	-33.73493	-25.64748	-25.26651	-16.24736	4.88061
β_{211}	0.97435	0.99421	0.99431	1.01396	0.01296
β_{212}	-0.02097	-0.00092	-0.00097	0.01937	0.01374
β_{220}	0.14591	7.85315	7.92533	16.27515	4.84033
β_{221}	-0.02845	-0.00759	-0.00748	0.01293	0.01371
β_{222}	0.96065	0.98355	0.98328	1.00375	0.01414
t_{211}	21.21069	50.69426	59.14515	145.98051	31.05766
t_{222}	21.21852	52.54299	61.03950	142.86673	36.27523
t_{221}	-61.15067	-4.91423	-7.03907	32.46757	23.00002
ϕ_{211}	14.91414	294.00309	343.20092	987.49207	327.03972
ϕ_{212}	8.54822	219.71499	274.40208	841.34769	279.74376
$\sigma^2_{\varepsilon_2 1}$	27.41238	52.05636	57.38080	118.65918	26.83870
$\sigma^2_{\varepsilon_2 2}$	32.18997	64.31691	69.13918	136.93742	31.40737

TABLE 3: StreetMap USA map model ($i = 3$) posterior summary.

Name	2.5%	50%	Mean	97.5%	IQR
β_{310}	-36.82311	-27.73408	-27.81309	-18.22047	5.31887
β_{311}	0.97524	0.99726	0.99693	1.01768	0.01456
β_{312}	-0.02617	-0.00313	-0.00307	0.02057	0.01458
β_{320}	-2.30644	5.65140	5.78789	14.70039	4.64570
β_{321}	-0.01990	-0.00095	-0.00115	0.01746	0.01247
β_{322}	0.96979	0.98933	0.98918	1.00974	0.01353
t_{311}	24.72349	61.81600	71.04935	181.72174	41.99450
t_{322}	19.56648	47.90725	54.44182	131.17618	32.07139
t_{321}	-59.21178	-5.46084	-6.20454	39.59914	22.79119
ϕ_{311}	22.07114	301.37057	339.04532	956.27422	301.73922
ϕ_{312}	6.79166	244.36028	303.00395	864.45841	315.41917
$\sigma^2_{\varepsilon_3 1}$	30.39977	61.59778	66.46791	127.44995	32.52752
$\sigma^2_{\varepsilon_3 2}$	27.63693	54.83787	59.28992	117.12683	27.64765

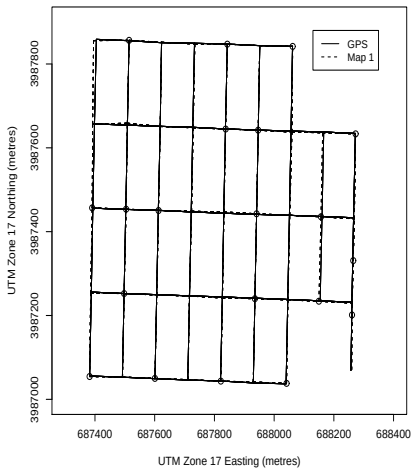

FIGURE 2: GPS-based map as in Figure 1 with map formed by posterior point estimates for map 1 using the relationship in (4). 20 circles centred on the point estimates illustrate approximate pinning corresponding to GPS points.

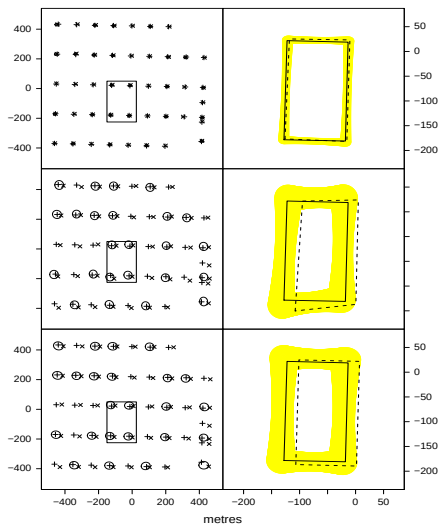

FIGURE 3: Left column: 95% credible ellipses based on normal approximation to posterior predictive draws of true locations for maps 1–3, top to bottom; + indicates GPS location, × map location. Right: corresponding 95% credible region for selected block indicated by enclosing rectangle at left. Solid rectangle: GPS. Dashed: map.

The subrectangle in each of the three plots on the left reveals the block feature we seek to predict. It is blown up in the respective right columns to show the 95% credible region for the true block. The region for the block was produced from approximate 95% credible ellipses for each edge point, say x, defined as a linear combination of the two vertices, say v_1 and v_2, that define the edge: $x = pv_1 + (1-p)v_2$, where $0 \leq p \leq 1$. These results attempt to capture the uncertainty in predicting the true block location and, again, illustrate the relative superior positional accuracy of the DLG ($i = 1$) map. Note the shape of the credible regions for lines connecting two vertices:

the regions are more narrow near the middle of the line (Figure 3, right); this shape is similar to existing characterisations of uncertainty for linear features, such as the modified epsilon-band (Caspary & Scheuring 1993; Leung & Yan 1998; Shi 1998; Shi & Liu 2000) and the G-band of Shi & Liu (2000).

While a map may have poor point positional accuracy, this need not imply correspondingly poor accuracy for relative measures like areas and distances. To this end, Table 4 gives predictive summaries of the area for the polygons shown in the right column of Figure 3. Table 4 summarises the predicted length of the top edge of these polygons. Values based on the GPS control points and each set of corresponding map points are given in the captions of the tables. Thus, we see that the absolute positioning of a map may be poor in the sense that true point credible ellipses may not contain corresponding map points, but that relative measures like area and distance may be more accurate in the sense that the credible intervals for these quantities do contain their corresponding map quantities, with, again, much higher uncertainty for maps 2 and 3.

TABLE 4: Posterior area (square metres) summary of the polygons shown in the right column of Figure 3. Observed values: 21887.98(GPS), 22249.31($i = 1$), 22454.71($i = 2$), 22910.02($i = 3$).

Map	2.5%	50%	Mean	97.5%	IQR
1	21194	22356	22350	23638	755
2	18069	21828	21804	25286	2296
3	19427	22735	22795	26397	2215

TABLE 5: Posterior linear distance (metres) summary of the top edge of the polygons shown in the right column of Figure 3. Observed values: 109.0402(GPS), 108.0185($i = 1$), 99.71289($i = 2$), 108.3128($i = 3$).

Map	2.5%	50%	Mean	97.5%	IQR
1	101.8	108.4	108.5	115.4	4.0
2	79.8	99.8	99.9	118.9	11.1
3	88.6	109.1	109.3	130.0	12.8

6. DISCUSSION

We have developed a flexible model for map positional error capturing global transformation, rubber-sheeting and GPS measurement error. We have described how available information on map accuracy can be used to provide useful prior information on the associated variance components. In the case of multiple maps, we have shown how to use Bayesian model averaging to predict true location using all of the maps. We have illustrated prediction not only with point features such as road intersections but also linear features such as roads and city blocks.

Our model is relatively straightforward to fit since it is built from bivariate Gaussian processes. We could readily enrich the model to include non-Gaussian error components and to allow nonstationarity in our spatial process modelling. However, the existing literature, when it does describe uncertainty, almost always does it through normality. Also while nonstationarity will often prove important in modelling responses at locations, it is arguably less of an issue for map positional error.

We note that our model development is based on the assumption of a common reference system. The use of the affine transformation would most certainly be inadequate to capture much

of the systematic error that exists between different coordinate systems. Still the availability of spatial databases and the ease with which they can be transformed to a common reference system makes the model widely applicable. Moreover, in principle, we could consider a more complex, perhaps nonlinear, large-scale transformation if we wish to consider the different sorts of systematic errors that may arise between different systems. However, we would caution against the use of our approach for maps with differing reference systems or for maps whose projection and datum information is unknown.

The primary use for our work is anticipated to involve taking data from a poorer quality map and introducing it onto a higher quality map where it is absent. In this setting, the higher quality map plays the role of the control data. For instance, how would we impute a feature on a TIGER data file to a DLG data file in which it does not appear and characterise the uncertainty in the imputed location? In this spirit, future work will investigate reconstruction problems: what would the true current location be of features found on historical maps, features that are no longer in existence? We also will examine the true location of more general curvilinear features such as municipal boundaries or shorelines.

ACKNOWLEDGEMENTS

This work was supported in part by a National Science Foundation grant. The authors thank Tom Meyer for valuable comments. They also thank the Associate Editor and two referees for helpful suggestions for improving the clarity of the presentation.

REFERENCES

P. M. Atkinson (1999). Geographical information science: geostatistics and uncertainty. *Progress in Physical Geography*, 23, 134–142.

S. Banerjee, B. P. Carlin & A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied probability 101, Chapman & Hall/CRC, Bota Raton.

M. Blakemore (1984). Generalisation and error in spatial data bases. *Cartographica*, 21, 131–139.

S. P. Brooks & A. Gelman (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.

R. J. Carroll, D. Ruppert & L. A. Stefanski (1995). *Measurement Error in Nonlinear Models*. Monographs on Statistics and Applied Probability 63, Chapman & Hall, London.

W. Caspary & R. Scheuring (1993). Positional accuracy in spatial data bases. *Computational, Environmental, and Urban Systems*, 17, 103–110.

S. Chib (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.

S. Chib & I. Jeliazkov (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96, 270–281.

N. Chrisman (1982). A theory of cartographic error and its measurement in digital databases. In *ISPRS 4 Proceedings, Auto-Carto 5: Fifth International Symposium on Computer-Assisted Cartography and International Society for Photogrammetry and Remote Sensing* (J. Foreman, ed.), American Society of Photogrammetry, Falls Church, pp. 159–168.

N. Cressie (1993). *Statistics for Spatial Data*, Revised edition. Wiley, New York.

N. Cressie & J. Kornak (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, 18, 436–456.

I. Dowman (1999). Encoding and validating data from maps and images. In *Geographical Information Systems: Principles, Techniques, Applications, and Management, Volume 2: Management Issues and Applications* (P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, eds.), Second edition, Wiley, New York, pp. 437–450.

J. Drummond (1995). Positional accuracy. In *Elements of Spatial Data Quality* (S. C. Guptill & J. L. Morrison, eds.), Elsevier Science, New York, pp. 31–58.

Environmental Systems Research Institute (2003). *ESRI Data and Maps ArcGIS (v8.2) StreetMap USA Local Streets CD-ROM 8*. Environmental Systems Research Institute, Redlands, CA.

W. A. Fuller (1987). *Measurement Error Models*. Wiley, New York.

J. Gabrosek & N. Cressie (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, 34, 262–285.

A. E. Gelfand & B. Mallick (1996). Semiparametric errors in variables models; a Bayesian approach. *Journal of Statistical Planning and Inference*, 52, 307–321.

A. Gelman, J. B. Carlin, H. S. Stern & D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton.

A. Gelman & D. B. Rubin (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–511

W. R. Gilks, S. Richardson & D. J. Spiegelhalter, eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Boca Raton.

M. F. Goodchild & S. Gopal, eds. (1989). *The Accuracy of Spatial Databases*. Taylor & Francis, London.

S. C. Guptill & J. L. Morrison, eds. (1995). *Elements of Spatial Data Quality*. Elsevier Science, New York.

G. B. Heuvelink (1999). Propagation of error in spatial modelling with GIS. In *Geographical Information Systems: Principles, Techniques, Applications, and Management, Volume 2: Management Issues and Applications* (P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, eds.), Second edition, Wiley, New York, pp. 207–217.

G. B. Heuvelink, P. A. Burrough & A. Stein (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Science*, 3, 303–322.

G. J. Hunter & M. F. Goodchild (1996). A new model for handling vector data uncertainty in geographic information systems. *URISA Journal*, 8, 51–57.

H. Kiiveri (1997). Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, 11, 33–52.

Y. Leung & J. Yan (1998). A locational error model for spatial features. *International Journal of Geographical Information Science*, 12, 607–620.

P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, eds. (1999a). *Geographical Information Systems: Principles, Techniques, Applications, and Management, Volume 1: Principles and Technical Issues*, Second edition. Wiley, New York.

P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, eds. (1999b). *Geographical Information Systems: Principles, Techniques, Applications, and Management, Volume 2: Management Issues and Applications*, Second edition. Wiley, New York.

K. Lowell & A. Jaton (Eds.) (1999). *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*. Ann Arbor Press, Chelsea, Michigan.

H. T Mowrer & R. G. Congalton (Eds.) (1999). *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*. Ann Arbor Press, Chelsea, Michigan.

J. Perkal (1966). On the length of empirical curves. In *Discussion Paper Number 10* (J. D. Nystuen, ed.), Michigan Inter-University Community of Mathematical Geographers, Ann Arbor, pp. 23–57.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna: http://www.R-project.org.

A. E. Raftery, D. Madigan & J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179–191.

W. Shi (1998). A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographical Information Science*, 12, 131–143.

W. Shi, P. F. Fisher & M. F. Goodchild (Eds.) (2002). *Spatial Data Quality*. Taylor & Francis, New York.

W. Shi & W. Liu (2000). A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science*, 14, 51–66.

T. Soler & J. Marshall (2002). Rigorous transformation of variance-covariance matrices of GPS-derived coordinates and velocities. *GPS Solutions*, 6, 76–90.

L. V. Stanislawski, B. A. Dewitt & R. L. Shrestha (1996). Estimating positional accuracy of data layers within a GIS through error propagation. *Photogrammetric Engineering & Remote Sensing*, 62, 429–433.

K. Thapa & J. Bossler (1992). Accuracy of spatial data used in geographic information systems. *Photogrammetric Engineering & Remote Sensing*, 58, 835–841.

Trimble Navigation (1997a). *TDC1 Asset Surveyor Operation Manual (31174-20)*. Sunnyvale, California.

Trimble Navigation (1997b). *TDC1 Asset Surveyor Software User Guide (31173-20)*. Sunnyvale, California.

Trimble Navigation (2003). *GPS Pathfinder Office Getting Started Guide (34231-30-ENG)*. Westminster, CO.

H. Veregin (1999). Data quality parameters. In *Geographical Information Systems: Principles, Techniques, Applications, and Management, Volume 1: Principles and Technical Issues* (P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, eds.), Second edition, Wiley, New York, pp. 177–189.

H. Wackernagel (2003). *Multivariate Geostatistics: An Introduction with Applications*, Third edition. Springer-Verlag, Berlin.

M. S. White, Jr. & P. Griffin (1985). Piecewise linear rubber-sheet map transformation. *The American Cartographer*, 12, 123–131.

Received 26 May 2005
Accepted 2 June 2006

Jarrett J. BARBER: jbarber8@uwyo.edu
Department of Statistics, University of Wyoming
Laramie, Wyoming 82071, USA

Alan E. GELFAND: alan@stat.duke.edu
Institute of Statistics and Decision Sciences
Durham, North Carolina 27708-0251, USA

John A. SILANDER, Jr.: john.silander@uconn.edu
Ecology and Evolutionary Biology Department
Storrs, Connecticut 06269-3043, USA