Bayesian Analysis (2008) **3**, Number 3, pp. 625–658

Combining multiple maps of line features to infer true position

Jarrett J. Barber* and Steven D. Prager†

Abstract. Map positional error refers to the difference between a feature's coordinate pair on a map and the corresponding true, unknown coordinate pair. In a geographic information system (GIS), this error is propagated through all operations that are functions of position, so that lengths, areas, etc., are uncertain. Often, a map's metadata provides a nominal statement on the positional error of a map, and such information has frequently been used to study the propagation of error through such operations. This article presents a statistical model for map positional error, incorporating positional error metadata as prior information, along with map coordinates, and, in particular, the information contained in the linearity of features. We demonstrate that information in the linearity of features can greatly improve the precision of true location predictions.

Keywords: GIS, linear features, lines, maps, positional error

1 Introduction

1.1 Positional error problem

The development of automated mapping, geographic information systems (GIS), remote sensing technology, and various other automated data collection and spatially aware communication devices has allowed us to amass and readily access an enormous amount of spatially referenced data. Our data collection proficiency has been accompanied by the concomitant development of a vast enterprise of management and science of geographic information (Longley et al. 1999b,a). Even so, the subject of spatial data quality—specifically, the uncertainty associated with positional information—is a relatively small part of this enterprise. At the same time, in spite of numerous approaches to understanding the role of uncertainty in affecting the quality of geographic information, there remains a lack of understanding with regard to best practices for reconciling positional information from multiple sources (Goodchild and Gopal 1989; Thapa and Bossler 1992; Guptill and Morrison 1995; Veregin 1999; Lowell and Jaton 1999; Mowrer and Congalton 1999; Shi et al. 2002; Zhang and Goodchild 2002). Given the increasing prevalence of geographic information originating from multiple sources and the commensurate lack of robust techniques for reconciling such information, new approaches are required. In this paper we thus focus on characterizing the uncertainty associated with feature positions reported from multiple data sources.

Many of the more statistically refined approaches to assessing spatial data uncertainty are based primarily on geostatistical methods applied to measured attributes recorded at locations, which are nearly always assumed to be known (Cressie 1993; Atkinson 1999; Banerjee et al. 2004). In the geographic information science (GIScience) literature, Goodchild (2004) and Le-

*Department of Statistics, University of Wyoming, Laramie, WY, mailto:jbarber8@uwyo.edu
†Department of Geography, University of Wyoming, Laramie, WY, mailto:SDPrager@uwyo.edu

 DOI:10.1214/08-BA325

ung, Ma and Goodchild (2004a; 2004b; 2004c; 2004d) contribute recent work on positional error modeling: borrowing from the map survey community (Wolf and Ghilani 1997), their "approximate law of error propagation" is an application of Taylor series approximation to (non–linear) transformations of positions, i.e., the delta method (Casella and Berger 2002). Even so, the majority of the work on positional error in the GIScience community does not occur within a statistical inference framework, and little effort has been directed toward estimation of positional error models or toward the prediction of true position using positional data and ancillary information. In the few instances within the statistical literature where positional error is discussed, the effort generally is to assess the effect on spatial prediction (Gabrosek and Cressie 2003; Cressie and Kornak 2003) or on attribute classification (Arbia et al. 2003). Only recently has the problem of positional error and prediction of true location caught the attention of the statistical community (Barber et al. 2006; Barber 2007).

Our approach to reconciling positional uncertainty across multiple "maps," i.e., geographic data layers, represents a substantial and novel step toward understanding how to use information in multiple, potentially disparate sources of positional information. In particular, our models of positional error that we present in this paper account for the linearity of map features, resulting in remarkably improved inference for true position compared to models that ignore feature linearity. The incorporation of map metadata into our models is important and calls naturally for a Bayesian approach.

In the process, we extend the earlier notion of the epsilon–band characterization of positional uncertainty in line segments (Perkal 1966; Chrisman 1982; Blakemore 1984; Veregin 1999) as well as the more recent, closely related refinements by various authors in the GIScience community (Caspary and Scheuring 1993; Leung and Yan 1998; Shi 1998; Shi and Liu 2000; Leung et al. 2004a). Our approach differs substantially from that of Barber et al. (2006) and does not require their model averaging (Raftery et al. 1997) approach. Our work may be viewed simply as an improved basis for inference of true position or functions thereof; as progress toward a fully model–based development of error propagation studies (Heuvelink et al. 1989; Heuvelink 1999); or as a precursor to future work that synthesizes geographic information theory with varied sources of data and information for improved inference of true position.

1.2 The modeling problem

Barber et al. (2006) provide a discussion of various potential approaches to modeling positional error, defined as—up to a change in sign—the difference between observed position on a map and unobserved true position. These modeling approaches adopt a GIS map registration perspective whereby one source of positional information is viewed as the best map or as the "target", e.g., GPS ground truth coordinates, toward which other sources are to be adjusted in some way. In our current effort, we depart from the map registration perspective. GPS locations or control points are viewed and modeled in a manner analogous to other types of positional data so that both GPS and other types of observed locations may simply be considered as different sources of positional information.

We view observed location as arising from and varying about latent true location, and we model each source of positional data as arising conditionally independent of other sources given the truth and possibly given other model parameters. Generically, for an observed coordinate pair $\boldsymbol{y}$ and corresponding unobserved true coordinate pair $\boldsymbol{x}$, we have $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$ where $\boldsymbol{e}$ may be thought of as a positional error vector (Figure 1). This approach may be viewed as a measurement–error model in the response (Carroll et al. 1995; Fuller 1987) as opposed to a

Berkson error model, wherein, instead, truth varies about observed. Barber et al. (2006) take the latter approach and offer more discussion on modeling approaches in a map context. The latent–truth approach is consistent with the measurement error (ME) model advocated by Goodchild (2004) and Leung, Ma and Goodchild (2004a; 2004b; 2004c; 2004d) in the context of measurement–based GIS (MBGIS) (Goodchild 1999) and is related to their approximate law of error propagation through transformations of position. However, they do not develop the specification of prior information, they do not consider the synthesis of multiple sources of positional data, and they do not incorporate feature linearity. We do.

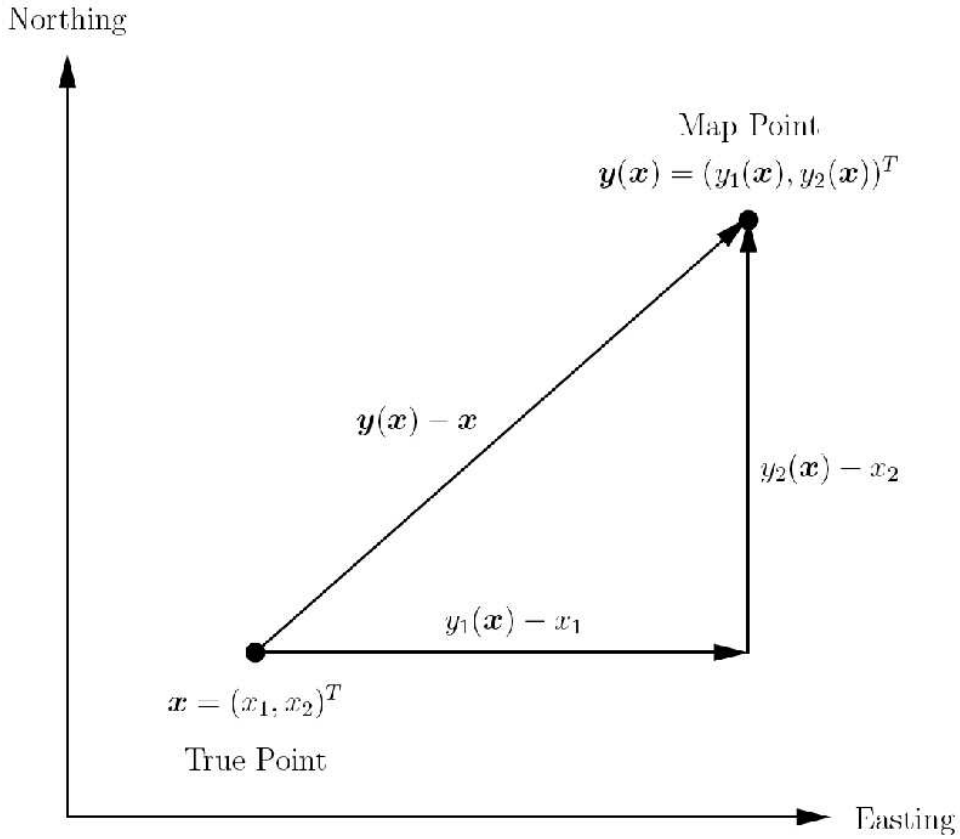

Figure 1: Suggestive diagram of positional error.

A process model for linear features

A simple but novel and remarkable development in this paper is the improvement of inference for true position obtained by incorporating feature linearity into a model of position. In particular, we model latent true coordinates as arising from one or more latent true lines as we might do for points along roads that are, practically speaking, straight, and that may be intersected by other straight line features. Feature linearity may be considered as a simple but

informative latent process. In the instance of road networks, the underlying line model may be considered to represent the process of designing, surveying, and constructing straight roads.

We have presented the notion that observed map points arise from underlying latent true points. If two map points are observed to be connected by a line segment, then there is the corresponding notion of an underlying true line segment. A collection of successive, observed line segments may depict a *curvilinear* feature, e.g., a stream or a beltway, conjuring the notion of a corresponding true curvilinear feature. We focus here on the special case of a *linear* feature comprised of a related set of connected segments that are considered collectively to depict a straight feature. Just as observed points are not true points, observed segments are not true segments, and we cannot expect the observed segments to lie perfectly along a line, though we may believe that the observed segments are intended collectively to depict an essentially straight feature. We associate each true linear feature with the equation of a line, and, throughout, we use "linear feature" and "line" interchangeably, and risk confusion that might arise from, say, the possibility that different linear features might fall on the same line or that linear features may not intersect while their associated lines do.

We note that it is not necessary to consider explicitly a line specification for a linear feature consisting of a single line segment since such a feature is determined by the segment's endpoints, which will have their own model specification. But, a linear feature depicted by two or more segments—associated with three or more points—is more than a set of connected points, and the points must be associated explicitly with a line equation if we are to gain information from the knowledge that the feature on which the points lie is truely straight.

Above, we alluded to the first, or data, stage of our model wherein observed coordinates vary about truth; we give more details in Section 2.3. In Section 2.4, we discuss details of a model for true coordinates and true linear features or, more precisely, for true line intercept and slope parameters. The model for true coordinates and lines may be considered as the second stage in a hierarchical development, with the particular network of intersecting lines and points thereon dictating stochastic and deterministic model components. See Section 2.4 and Appendix 4. Also in Section 2.4, we will consider an exchangeable model structure for line slopes at a third stage for the case when line features occur in a grid, as in an orthogonal street network. We illustrate our approach using several sources of positional information for a residential neighborhood in Durham, North Carolina, USA (Section 3) and conclude with a discussion of our approach (Section 4).

2 Model

2.1 Notation

Consider M different sources of map information or other positional data. In the following development, it is not necessary to assume that maps depict the same region, although, in practice, maps will typically be associated with a common region of interest $\mathcal{X}$ that, for this article, will be taken to be a subset of a 2-dimensional, orthogonal coordinate system. For each coordinate pair on each data source, we consider a corresponding unobserved true coordinate pair. In general, there exists a one–to–many relationship between true coordinates and map coordinates as more than one map may depict the same feature. For example, a road intersection may be reported with different coordinate pairs on different maps, but these pairs correspond to the same true coordinates of the intersection. We work with $M = 4$ different

sources of positional data and denote each source with $m \in \{1, 2, 3, 4\}$ or $m \in \{C, D, G, S\}$, where the latter designation is suggestive of the particular sources to be described in Sections 2.2 and 3.

Let $\boldsymbol{x} \equiv (x_1, x_2)^T \in \mathcal{X}$ denote a generic true coordinate pair where x_1 and x_2 may be thought of, respectively, as the east–west (EW) coordinate and the north–south (NS) coordinate. Throughout, we use subscript $q = 1, 2$ to denote the components of $\boldsymbol{x}$, similarly for other coordinate pair vectors introduced below. Following the first–stage modeling introduction in Section 1.2 and using function notation, we denote $\boldsymbol{y}_m(\boldsymbol{x}) \equiv (y_{m1}(\boldsymbol{x}), y_{m2}(\boldsymbol{x}))^T$ as the corresponding generic coordinate pair on a map $m = 1, \ldots, M$. In our development, $\boldsymbol{y}_m(\boldsymbol{x})$ is observed for at least one map but may be missing for some maps. Assuming $\boldsymbol{x}$ and $\boldsymbol{y}_m(\boldsymbol{x})$ are in a common reference system (e.g., WGS 84 UTM meters), we define for map m the true positional error vector, $\boldsymbol{y}_m(\boldsymbol{x}) - \boldsymbol{x}$. Upon translation, the resulting vector may be viewed as a random vector emanating from truth and extending to observed (Figure 1).

Though a map depicts an uncountable set of locations, the representation of positional information associated with discrete geographic features is comprised of a finite set of coordinate pairs or "points". Hence, we are led to consider the j^{th} true point, $\boldsymbol{x}_j \equiv (x_{j1}, x_{j2})^T$, $j = 1, \ldots, N < \infty$, associated with a point observed on map m for at least one m. That is, the set of N true points under consideration is the union of all true points represented by observed points on one or more of the M maps. The number of points observed on map m may be less than N, and the set of map points $\{\boldsymbol{y}_m(\boldsymbol{x}_j) = (y_{m1}(\boldsymbol{x}_j), y_{m2}(\boldsymbol{x}_j))^T : j = 1, \ldots, N\}$ may include points that are missing on map m but observed on map m' for at least one $m' \neq m$. With occasion to distinguish observed from unobserved map points, we let $\{j_m(k) : k = 1, \ldots, n_m \leq N\}$ be a subset of $\{j : j = 1, \ldots, N\}$ such that $\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)}) = (y_{m1}(\boldsymbol{x}_{j_m(k)}), y_{m2}(\boldsymbol{x}_{j_m(k)}))^T$ is observed on map m.

As mentioned in Section 1.2, we also explicitly consider latent true lines, $l = 1, \ldots, L$, which are associated with a linear feature defined by three or more true points. In principle and without loss of generality, we may consider each of these lines parameterized as $x_2 = a_l + b_l x_1$. In Section 4, we discuss briefly implementation issues that may arise regarding line parameterization.

All distributions herein follow parameterizations in Gelman et al. (2003).

2.2 Prior information

In our application, Section 3, we combine four sources of positional data into a model to infer true position. Here, we discuss prior information on postitional error that we use for prior model specifications, which we discuss in Sections 2.3 and 3.1.

Three sources are extracted from data layers stored within a GIS. Such data layers are typically accompanied by metadata that can be used to specify prior distributions for the positional error variance components of the corresponding layer. For example, we use a large–scale (1:24000) USGS Digital Line Graph (DLG) (U.S. Geological Survey 2004, 1999) ($m = D$) which is accompanied by one such measure of positional error, root mean square error (RMSE). In particular, locations on large–scale Digital Line Graphs (DLGs) produced by the US Geological Survey (USGS)

> *...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,*

the "standard error" simply being the square root of the mean squared differences between coordinates—EW or NS—on the DLG and the corresponding source map coordinates (U.S. Geological Survey 1999, Sections 2.3.4 and 2.4.2).

The source maps of most large–scale DLGs are USGS 7.5 minute topographic quadrangle maps with a map scale of 1:24000. These quadrangle maps are produced 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* (U.S. Bureau of the Budget 1947).

At a map scale of 1:24000, this means that 90% of tested map points fall within 40 ft of corresponding reference points that are effectively considered "truth." If we assume a bivariate Gaussian distribution for the positional error vector having uncorrelated components with common scale, this corresponds to the 90% error circle of $\mathrm{N}(\mathbf{0}, (18.64\,\mathrm{ft})^2\boldsymbol{I}_2)$, $\boldsymbol{I}_2$ being the 2×2 identity matrix. At the same map scale, the DLG production process discussed above nominally introduces a standard error less than 6 ft. Thus, if we assume positional error vector components to be uncorrelated normal variates with common variance, and if we assume the DLG production process to be independent of the production of quadrangle maps, then we should conclude that a large–scale DLG has total *a priori* positional error variance less than about $20^2\ \mathrm{ft}^2$ ($18.64^2 + 6^2$) or about $6^2\ \mathrm{m}^2$. We denote this value as $\sigma^2_{t,m}$, $m \in \{C, D, G, S\}$, e.g., $\sigma^2_{t,D} = 6^2\ \mathrm{m}^2$ for the DLG. We consider $\sigma^2_{t,m}$ to refer to the positional error vector, $\boldsymbol{y}_m(\boldsymbol{x}) - \boldsymbol{x}$. That is, $\sigma^2_{t,m}$ is considered an *a priori* value for the diagonals of $\mathrm{Var}(\boldsymbol{y}_m(\boldsymbol{x}) - \boldsymbol{x} \,|\, \boldsymbol{x})$. We qualify $\boldsymbol{y}_m(\boldsymbol{x}) - \boldsymbol{x}$ as the *total* positional error vector to distinguish it from another error vector arising from subsequent modeling efforts. Hence, we refer to $\sigma^2_{t,m}$ as the *a priori total* positional error variance; we offer further discussion in Section 2.3. Prior information for two other sources of positional data used here may be specified in a manner similar to that discussed for the DLG.

The remaining, fourth set of coordinate pairs used here was obtained by means of a global positional system (GPS) ($m = G$). Each of these GPS points is, more precisely, an average of approximately 200 differentially corrected GPS position fixes. It is straightforward to use directly the n_G sets of 200 corrected GPS fixes, and, in this case, these sets of points may be used to estimate a 2×2 covariance matrix for GPS (Soler and Marshall 2002), i.e., to estimate the entries in $\mathrm{Var}(\boldsymbol{y}_G(\boldsymbol{x}) - \boldsymbol{x} \,|\, \boldsymbol{x})$.

We choose instead to use the averages and, furthermore, to use additional, post–processing information (Trimble Navigation 1997a,b, 2003) to specify "prior" distributions for the GPS positional error variance components in $\mathrm{Var}(\boldsymbol{y}_G(\boldsymbol{x}) - \boldsymbol{x} \,|\, \boldsymbol{x})$. We take this approach because the averages typically are used in practice, not the raw fixes, and the additional post–processing information should, in good practice, accompany the point averages as metadata. Hence, we consider GPS post–processing information to be *a priori* information analogous to the previously discussed positional sources' metadata—though it should be clear that, for GPS, such information strictly may not be considered as "prior" information since it is obtained from the raw fixes via the GPS post–processing procedure.

This discussion of prior information is focused on the use of a positional data source's metadata to specify an informative prior distribution for variance parameters associated with the components of the positional error vector, and we utilize such information in our application. While we do not have prior information to help specify our line models—we use non–informative priors for slopes and intercepts—it seems reasonable that an expert knowledge of surveying

and constructing linear features, e.g., roads, may be useful for such specifications.

We give further details on priors in Sections 2.3 and 3.1.

2.3 Positional data source model

A point–wise specification

Following the discussion in Section 1.2, the model for the $j_m(k)^{th}$ observed point on source m is

$$\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)}) = \boldsymbol{\mu}_m(\boldsymbol{x}_{j_m(k)}; \boldsymbol{\beta}_m) + \boldsymbol{\varepsilon}_m(\boldsymbol{x}_{j_m(k)}),$$

where $\boldsymbol{\mu}_m(\boldsymbol{x}; \boldsymbol{\beta}_m)$ is some function of true location $\boldsymbol{x}$ and parameter $\boldsymbol{\beta}_m$, and, using function notation similar to $\boldsymbol{y}_m(\boldsymbol{x})$, $\boldsymbol{\varepsilon}_m(\boldsymbol{x})$ is a bivariate Gaussian error vector with mean $\mathbf{0}$. Though we can write, generically, $(\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{x}) = (\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}) + \boldsymbol{\varepsilon}(\boldsymbol{x})$, we resist referring to $\boldsymbol{\varepsilon}(\boldsymbol{x})$ as a positional error error or, even less attractively, as a positional error vector error vector! Obviously, if $\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{x}$, then $\boldsymbol{\varepsilon}(\boldsymbol{x})$ is a what we have referred to as a *total* positional error vector (Section 2.2). Otherwise, if $\boldsymbol{\mu}(\boldsymbol{x}) \neq \boldsymbol{x}$, then $\boldsymbol{\varepsilon}(\boldsymbol{x})$ is, strictly speaking, different. In any case, we see little harm in referring to $\boldsymbol{\varepsilon}(\boldsymbol{x})$ as error, positional error, or (residual) postional error vector, at least until it is necessary to use the *total* qualifier as we do in Section 2.3 below.

In our application, we consider the error vectors $\boldsymbol{\varepsilon}_m(\boldsymbol{x}_j)$ to be conditionally independent across $m = 1, \ldots, M$, $j = 1, \ldots, N$, and $q = 1, 2$, given the true points $\boldsymbol{x}_j$. In the following model development, however, we maintain some generality for the covariance structure of the $\boldsymbol{\varepsilon}_m(\boldsymbol{x})$, but maintain independence across m throughout. We define the conditional variance matrix for $\boldsymbol{\varepsilon}_m(\boldsymbol{x})$, given $\boldsymbol{x}$, to be

$$\boldsymbol{V}_m \equiv \begin{pmatrix} \sigma_{m1}^2 & \sigma_m \\ \sigma_m & \sigma_{m2}^2 \end{pmatrix} \equiv \mathrm{Var}(\boldsymbol{\varepsilon}_m(\boldsymbol{x})), \tag{A.1}$$

allowing for separate positional error vector component variances, σ_{m1}^2 and σ_{m2}^2, and covariance σ_m.

In principle, we may use any function $\boldsymbol{\mu}_m(\boldsymbol{x})$ for the mean of observed positions on source m. Here, we consider a six–parameter affine transformation, a common large–scale transformation in map registration procedures, that allows for shifting, scaling, and rotating of one set of points relative to another—observed relative to true in our case. See, e.g., Dowman (1999) or Barber et al. (2006). The affine transformation for $\boldsymbol{\mu}_m(\boldsymbol{x}) \equiv (\mu_{m1}(\boldsymbol{x}), \mu_{m2}(\boldsymbol{x}))^T$ is

$$\begin{aligned} \mu_{m1}(\boldsymbol{x}) &\equiv \beta_{m10} + \beta_{m11} x_1 + \beta_{m12} x_2 \\ \mu_{m2}(\boldsymbol{x}) &\equiv \beta_{m20} + \beta_{m21} x_1 + \beta_{m22} x_2. \end{aligned} \tag{A.2}$$

Using the point–wise probability specification,

$$p(\boldsymbol{y}_m(\boldsymbol{x})) = \mathrm{N}(\boldsymbol{\mu}_m(\boldsymbol{x}; \boldsymbol{\beta}_m), \boldsymbol{V}_m),$$

we may write the observed likelihood as

$$\prod_{m=1}^{M} \prod_{k=1}^{n_m} p(\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})), \tag{A.3}$$

where $\boldsymbol{\beta}_m \equiv (\beta_{m10}, \beta_{m20}, \beta_{m11}, \beta_{m21}, \beta_{m12}, \beta_{m22})^T$. For a conditionally independent coordinate–wise specification, that is, when, additionally, $\sigma_m = 0$, the analogous coordinate–wise expression is obvious, and we omit it.

The above specification results in a multivariate Gaussian distribution, jointly for all observed source coordinates conditional on true locations. We describe this distribution in more generality next. We actually implement the coordinate–wise independent specification and discuss priors in this context in Sections 2.3 and 3.1.

A matrix specification

Stack observed coordinate pairs to get

$$\tilde{\boldsymbol{y}}_m \equiv (\boldsymbol{y}_m^T(\boldsymbol{x}_{j_m(1)}), \ldots, \boldsymbol{y}_m^T(\boldsymbol{x}_{j_m(n_m)}))^T \tag{A.4}$$

for each m, and define $\boldsymbol{\Sigma}_m \equiv \mathrm{Cov}(\tilde{\boldsymbol{y}}_m) = \mathrm{Cov}(\tilde{\boldsymbol{\varepsilon}}_m)$, where $\tilde{\boldsymbol{\varepsilon}}_m$ is defined analogously to $\tilde{\boldsymbol{y}}_m$. Thus, $\boldsymbol{\Sigma}_m$ describes the covariance structure among the coordinates of positional source m. Again, for the moment, we consider a general covariance structure for $\boldsymbol{\Sigma}_m$ and return to this issue in the Discussion. Define the augmented matrix of true coordinates

$$\boldsymbol{X} \equiv \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{pmatrix},$$

and let $\boldsymbol{H}_m$ be the $n_m \times N$ indicator matrix with a one in the j^{th} column of row k if $j_m(k) = j$, zero otherwise. Thus, a 1 in element $(k, j_m(k))$ of $\boldsymbol{H}_m$ indicates that observed point $\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})$ on source m corresponds to true point $\boldsymbol{x}_{j_m(k)}$. Since $\{j_m(k) : k = 1, \ldots, n_m \leq N\}$ is a subset of $\{j : j = 1, \ldots, N\}$, then $\boldsymbol{H}_m$ has a single value of 1 in each row, zeros elsewhere. For example, if $n_m = 2$ and $N = 3$ with $j_m(1) = 1$ and $j_m(2) = 3$, then

$$\boldsymbol{H}_m = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Stacking consecutively the observation vectors (A.4) of each source, define $\tilde{\boldsymbol{y}} \equiv (\tilde{\boldsymbol{y}}_1^T, \ldots, \tilde{\boldsymbol{y}}_M^T)^T$, and, likewise, stack to get $\tilde{\boldsymbol{\varepsilon}} \equiv (\tilde{\boldsymbol{\varepsilon}}_1^T, \ldots, \tilde{\boldsymbol{\varepsilon}}_M^T)^T$ and $\tilde{\boldsymbol{\beta}} \equiv (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_M^T)^T$. Then, the likelihood for all M sources follows from

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \mathrm{Diag}(\boldsymbol{H}_1\boldsymbol{X} \otimes \boldsymbol{I}_2, \ldots, \boldsymbol{H}_M\boldsymbol{X} \otimes \boldsymbol{I}_2)\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}} \\ &\equiv \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}} \quad \text{and} \\ \tilde{\boldsymbol{\varepsilon}} &\sim \mathrm{N}\left(0, \mathrm{Diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)\right), \end{aligned}$$

where Diag returns a block–diagonal matrix.

From prior information to model specification

Here, we make a transition from prior information (Section 2.2) to the specification of prior distributions for the parameters of the coordinate–wise source model of Section 2.3. To facilitate the transition, we introduce, shortly, a sort of *a priori* source model analogous to that of Section 2.3, but whose development is motivated by the *a priori* considerations of Section 2.2. Both the

notions of accuracy and precision are relevant, and we discuss these briefly before specifying priors. Particular numerical values of prior distribution parameters are given in Section 3.1.

A map's observed points may be accurate in the sense that $\boldsymbol{\mu}_m(\boldsymbol{x}) = \boldsymbol{x}$, or inaccurate on the contrary. Either case may occur with either high or low precision as allowed for by the error variance components σ^2_{m1} and σ^2_{m2}. While a map may be (in)accurate in the above sense, it may retain *relative positional* accuracy in that, for example, distances, angles, and directions may be accurate, which may occur with high or low precision. We may refer to the former notion of accuracy as *absolute positional* accuracy. We do not see a clear distinction between absolute positional precision and relative positional precision. The typical "target analogy" of accuracy and precision is obvious.

The sources of positional data that we use here are either from a well–established production series of source maps or from a GPS (Section 2.2). In either case, the assumption of *a priori* absolute accuracy seems eminently reasonable. This assumption implies that our *a priori* model is $\boldsymbol{y}_m(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\varepsilon}_m(\boldsymbol{x})$, and this model seems to be implicit in metadata statements of "map accuracy," such as those given in Section 2.2, though, according to our usage, these statements seem to refer to *precision*, with *accuracy* implied. More precisely, we assume such statements refer to the *total* positional error vector $\boldsymbol{y}_m(\boldsymbol{x}) - \boldsymbol{x} = \boldsymbol{\varepsilon}_m(\boldsymbol{x})$, given $\boldsymbol{x}$, and are meant to refer to the entire production series—a sort of map series production standard.

In our experience, such *a priori* information rarely helps to draw a distinction between the components of the error vector or to suggest how they are related. Thus, the following *a priori* variance model seems reasonable:

$$\boldsymbol{y}_m(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\varepsilon}_m(\boldsymbol{x})$$

with

$$\boldsymbol{V}_m = \begin{pmatrix} \sigma^2_{t,m} & 0 \\ 0 & \sigma^2_{t,m} \end{pmatrix},$$

using the previously discussed *a priori total* positional error variance component $\sigma^2_{t,m}$ in place of σ^2_{m1} and σ^2_{m2} in (A.1). Assuming accuracy holds, then zero covariance is reasonable and is supported by results reported by Barber et al. (2006).

While the above discussion suggests that an entire production series is accurate in some overall, *a priori* sense, particular maps in a series often exhibit what may be characterized as systematic variation from (a source known to be much closer to) truth. This suggests the use of a mean other than $\boldsymbol{\mu}_m(\boldsymbol{x}) = \boldsymbol{x}$. Using the affine model (A.2) as an example, we note that $\boldsymbol{\beta}_m = (0, 0, 1, 0, 0, 1)^T$ corresponds to *a priori* accuracy. Of course, we specify a prior distribution for $\boldsymbol{\beta}_m$, and we acknowledge *a priori* accuracy by centering the prior at $\mathrm{E}(\boldsymbol{\beta}_m) = (0, 0, 1, 0, 0, 1)^T$. A flat, improper prior suggests that we have no *a priori* notion of accuracy, but an improper prior on $\boldsymbol{\beta}_m$ results in an improper posterior if true coordinates are specified stochastically, which they are (Section 2.4). In the interest of propriety, we choose independent *bounded* uniform priors for the elements of $\boldsymbol{\beta}_m$, $m \in \{C, D, S\}$, with prior mean $(0, 0, 1, 0, 0, 1)^T$.

However, for $\boldsymbol{\beta}_G$, we specify a point mass prior at $(0, 0, 1, 0, 0, 1)^T$. In other words, according to (A.2), $\boldsymbol{x}$ is the mean of $\boldsymbol{y}_G(\boldsymbol{x})$, i.e., we assume GPS to be accurate. Furthermore, we incorporate the precision of GPS points via informative priors on the σ^2_{Gq}, to be discussed shortly. These informative specifications of accuracy and precision are sufficient to alleviate non–identifiability that may otherwise arise in absense of other information. Indeed, the incorporation of informative specifications in a Bayesian framework is an important aspect of

our approach. Thus, in this article, GPS may be considered as the "gold standard" relative to the other sources used here. In other instances, a few highly precise and accurate survey measurements may serve as the gold standard.

Denoting the above marginal priors as $p(\beta_{mqr})$, $m = 1, \ldots, M$, $q = 1, 2$, $r = 0, 1, 2$, and borrowing the $\tilde{\boldsymbol{\beta}}$ notation from Section 2.3, we use $p(\tilde{\boldsymbol{\beta}}) = \prod_m \prod_q \prod_r p(\beta_{mqr})$ to denote the joint prior for all beta parameters.

With the beta parameters having the aforementioned *a priori* accuracy, then $\sigma^2_{m1} = \sigma^2_{m2} = \sigma^2_{t,m}$ corresponds to *a priori* precision, and we specify independent scaled inverse–$\chi^2(\nu_m, s^2_m)$ priors for the σ^2_{mq}, each centered at mean $\mathrm{E}(\sigma^2_{mq}) = \frac{\nu_m}{(\nu_m - 2)} s^2_m \stackrel{\text{set}}{=} \sigma^2_{t,m}$, $m \in \{C, D, G, S\}$, $q = 1, 2$, with degrees of freedom ν_m. Using $p(\sigma^2_{mq})$, $m = 1, \ldots, m$, $q = 1, 2$, to denote these independent marginals, we write the joint prior as $p(\boldsymbol{\sigma}^2) = \prod_m \prod_q p(\sigma^2_{mq})$, where $\boldsymbol{\sigma}^2$ collects all of the σ^2_{mq} into a vector.

2.4 True points and lines

Points

In the previous section, observed source points are modeled conditionally on true points. As alluded to in Section 1.2, true coordinates may be stochastic or may be determined, depending on the network of intersecting true lines and the points thereon. The specification for points is simple if we condition on lines.

First, consider a point that does not fall on a line. In this case, both point's coordinates receive a stochastic specification; there are no lines to determine such points. For a point $\boldsymbol{x}_j = (x_{j1}, x_{j2})^T$ falling on only one line, say line l, we have $x_{j2} = a_l + b_l x_{j1}$, so that the coordinate x_{j2} is determined but x_{j1} is "free" and requires a stochastic specification; we consider line parameter specification shortly. In the case that a point falls on two and only two lines, i.e., it is the point of intersection of two lines, say, without loss of generality, lines 1 and 2, then we have $x_{j2} = a_1 + b_1 x_{j1}$ and $x_{j2} = a_2 + b_l x_{j2}$, which leads to

$$\begin{aligned} x_{j1} &= \frac{a_2 - a_1}{b_1 - b_2} \quad \text{and} \\ x_{j2} &= \frac{a_2 b_1 - a_1 b_2}{b_1 - b_2}, \end{aligned}$$

thus determining both point's coordinates, given line parameters.

More generally, if a point falls on more than two lines, in principle we may choose any two of the lines comprising the intersection to determine the coordinates in the above manner. However, in some scenarios, it turns out that the ease of line model specification may depend on which two lines are chosen. See Section 2.4 and Appendix 4. It should be clear that a true map consisting of a regular grid of streets—that are modeled as straight—may have very few coordinates, perhaps none, that require stochastic specification. In our application, all but three points are determined as points of line intersections, and the remaining three points fall on only a single line so that only one coordinate for each of these points requires a stochastic specification; see Figure 2.

It would be convenient to assign a flat, improper prior to all such stochastic true coordinates, but this would result in an improper posterior when using a stochastic specification for $\tilde{\boldsymbol{\beta}}$, which we do. Instead, we assign all undetermined EW coordinates x_{j1}, $j = 1, \ldots, N$,

to have independent and identical uniform priors supported by an interval encompassing all plausible true EW coordinates associated with all maps under consideration, where plausibility is aided by the maps' *a priori* total positional error variances $\sigma^2_{t,m}$, $m \in \{C, D, G, S\}$. Similarly, we assign all undetermined NS coordinates x_{j2}, $j = 1, \ldots, N$ to have independent and identical uniform priors supported by an interval encompassing all plausible true NS coordinates. Collecting true coordinates in $\tilde{\boldsymbol{x}}$, analogous to $\tilde{\boldsymbol{y}}$, above, we denote the product of all of the independent uniform priors for true coordinates as $p(\tilde{\boldsymbol{x}})$, though in doing so we risk confusion with the fact that, when some points lie on one or more lines, not all true coordinates in $\tilde{\boldsymbol{x}}$ have a stochastic specification. In addition, for simplicity of notation, conditioning on lines is suppressed when points fall on lines.

In the case that points do not fall on lines or if lines are not modeled, we now have everything we need to specify a full probability model for all quantities. Using the above established prior notation and the point–wise likelihood (A.3), this points–only model is

$$\left(\prod_{m=1}^{M} \prod_{k=1}^{n_m} p(\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})) \right) p(\tilde{\boldsymbol{\beta}}) p(\boldsymbol{\sigma}^2) p(\tilde{\boldsymbol{x}}). \tag{A.5}$$

In Section 3, we conduct a comparison among models including this "points–only" model and the line models described and enumerated below.

Lines

The data, i.e., source, model of Section 2.3 is conditional on true coordinates, among other parameters, and its specification is straightforward. And, there is no difficulty with regard to model specification for true coordinates given true lines—stochastic or deterministic as the case may be. But, a course of action for a true line model is not immediately obvious in the general case. However, most map applications will consist of true points that are associated with at most two true lines. In this case, the specification for line parameters is simple: all line parameters receive a stochastic specification. Again, this is the case for our application.

Cases involving more than two lines intersecting at the same point may lead to complications with regard to line parameter specification and identifiability. Because most applications do not fit this scenario, including our own application, we do not treat this scenario systematically. However, we do present a toy problem that introduces complications that may arise in the more general case (Appendix 4). Appendix 4 may help also to elucidate the simpler cases.

Thus, in most cases, point coordinate and line parameter specification is straightforward, and we note the reduction in required stochastic specifications for point coordinates when lines are modeled. For example, in our application, we have $N = 44$ true points, 41 of which are determined by line intersections, with only one coordinate from each of the remaining three points requiring stochastic specification. Since we have $L = 14$ lines in a simple configuration, we require 28 stochastic specifications for the slopes and intercepts. Thus, a total of 31 true line or point quantities require stochastic specification compared to 88 in a points–only model.

Again, for simplicity in notation, we denote the models for line intercepts and slopes as $p(\boldsymbol{a})$ and $p(\boldsymbol{b})$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ collect intercepts and slopes, respectively, and we hide possible determinedness of some parameters, though none are determined in our application.

A natural vague prior for a slope is uniform$(-\pi/2, \pi/2)$, on the radian scale, which, by transformation, results in a standard Cauchy prior on the original slope "rise–over–run" scale.

Thus, assuming *a priori* independence, $p(\boldsymbol{b})$ may be specified as the product of Cauchy densities. For intercepts, we appeal to the point–slope form, $(x_{j2} - x_{02}^l) = b_l(x_{j1} - x_{01}^l)$, where (x_{01}^l, x_{02}^l) is the "point," $l = 1, \ldots, L$. We specify independent uniform priors for all such point coordinates x_{0q}^l $l = 1, \ldots, L$, $q = 1, 2$, whose independence from slopes seems natural. In the Application, Section 3, we give particular bounds on the prior support of these coordinates that are appropriate to the source maps used there.

This specification for the point–slope form implies, *a priori*, that lines occur randomly over a region containing the maps of interest. Note that the points in the point–slope form are not identifiable in the sense that a given line l passes through an uncountable number of such points that would suffice in the point–slope form. But, this is not a problem since any such point on line l, together with the slope, b_l, uniquely determines the intercept, $a_l = -b_l x_{01}^l + x_{02}^l$, $l = 1, \ldots, L$. We retain notationally $p(\boldsymbol{a})$ as the (induced) prior on intercepts.

Using the above model development and the point–wise likelihood (A.3), we can write a full–probability model for all observed and unobserved quantities as

$$\left(\prod_{m=1}^{M} \prod_{k=1}^{n_m} p(\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})) \right) p(\tilde{\boldsymbol{\beta}}) p(\boldsymbol{\sigma}^2) p(\tilde{\boldsymbol{x}}) p(\boldsymbol{a}) p(\boldsymbol{b}), \tag{A.6}$$

and, again, we remind the reader that our notation favors brevity at risk of hiding the possible determinedness of some true coordinates and of some line parameters. In the application of this line model (A.6) in Section 3, none of the lines' parameters are determined, and all but three coordinates are determined. This is true also for the other line models described below.

So far, we have the "points–only" model (A.5) and refer to the current line model (A.6) as the "lines model". We introduce the "parallel lines" model (A.7) and "orthogonal lines" model (A.8) below.

Borrowing strength. Here, we investigate the cases of parallel lines and orthogonal lines. Figure 2 in Section 3 suggests that streets appearing to run nearly NS were intended to have the same slope and similarly for the streets that appear to run nearly EW. In this case, we replace the model, $p(\boldsymbol{b})$, in the lines model (A.6) with one of two other model specifications, which we now describe.

In general, there may be several sets of lines, each line in a set sharing approximately the same slope as the other lines in the same set. For example, Figure 2 suggests an EW set and an NS set. There are $L = 14$ line features—streets—indexed by $l \in \{1, 2, \ldots, 14\}$, which we divide into an EW set, $\{l_{EW}(k) : k = 1, \ldots, L_{EW} \leq L\} = \{1, 2, 3, 4, 5\}$, and an NS set, $\{l_{NS}(k) : k = 1, \ldots, L_{NS} \leq L\} = \{6, 7, \ldots, 14\}$.

In the lines model (A.6), we indicated independent standard Cauchy priors for the slopes, uniforms on the radian scale. Here, to borrow strength amongst slopes in the same set, we specify independent Cauchy priors for the means, μ_{EW} and μ_{NS}, of two parent distributions, then specify $b_{l_{EW}(k)} \overset{\text{iid}}{\sim} N(\mu_{EW}, \sigma_{EW}^2)$, $k = 1, \ldots, L_{EW}$, and $b_{l_{NS}(k)} \overset{\text{iid}}{\sim} N(\mu_{NS}, \sigma_{NS}^2)$, $k = 1, \ldots, L_{NS}$, where $\sigma_{EW}^2 = (1 + \mu_{EW})^2 \sigma_b^2$ and $\sigma_{NS}^2 = (1 + \mu_{NS})^2 \sigma_b^2$. Modeling the variance in this fashion gives an approximately contant variance on the radian scale by way of a first order Taylor approximation. That is, if $b \sim N(\mu, (1 + \mu)^2 \sigma^2)$, and $\theta = g(b) = \arctan(b)$, then $\text{Var}(\theta) \approx \sigma_b^2$.

It should be clear that it is on the radian scale that we should expect constant variance, not on the rise–over–run scale; a change of one unit for b centered at zero would tend to be

a much larger change in angle than a change of one unit for b centered at 100, but, certainly, we should not expect that the precision of the process of constructing straight streets depends on direction in this way! We could have given each parent slope distribution its own, separate variance, but, with a relatively small number of lines in each set, we chose to use all lines to estimate σ_b^2 in the weighted variance terms of both parents.

The common scale parameter, σ_b, may be assigned an improper distribution $p(\sigma_b) \propto \text{uniform}(0, \infty)$, giving a folded–noncentral–t full–conditional (Gelman 2006), but we specify, instead, a relatively large finite upper bound, and use WinBUGS (Lunn et al. 2000) for sampling.

Using, by now, familiar notation, we may write the parallel lines model as

$$\left(\prod_{m=1}^{M} \prod_{k=1}^{n_m} p(\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})) \right) p(\tilde{\boldsymbol{\beta}}) p(\boldsymbol{\sigma}^2) p(\tilde{\boldsymbol{x}}) p(\boldsymbol{a}) p(\boldsymbol{b}_{NS}) p(\boldsymbol{b}_{EW}) p(\mu_{NS}) p(\mu_{EW}) p(\sigma_b), \tag{A.7}$$

and, once again, we remind the reader that $p(\boldsymbol{b}_{NS})$ and $p(\boldsymbol{b}_{EW})$ may hide determineness, similar to our previous notation.

In an obvious extension, we consider an orthogonal street network or "grid," wherein lines share approximately the same slope, or slopes are approximately negative inverses of each other. In this case, we specify slopes or negative inverse slopes as arising independently from a single Gaussian parent, $N(\mu_O, \sigma_O^2)$, and denote the resulting joint distribution as $p(\boldsymbol{b}_O)$ where $\boldsymbol{b}_O$ collects slopes and negative inverse slopes. In Section 3, $\boldsymbol{b}_O$ consists of the slopes of near–EW lines and the negative inverse slopes of near–NS lines. Though we have no problem with more general cases—using some combination of model components already presented—we assume, for simplicity, that *all* slopes or negative inverse slopes belong to this one parent, which is the case in our application, and we use a standard Cauchy prior, $p(\mu_O)$, for the mean and assign a bounded uniform prior, $p(\sigma_O)$, to the standard deviation. The full–probability distribution for this orthogonal lines model may now be written as

$$\left(\prod_{m=1}^{M} \prod_{k=1}^{n_m} p(\boldsymbol{y}_m(\boldsymbol{x}_{j_m(k)})) \right) p(\tilde{\boldsymbol{\beta}}) p(\boldsymbol{\sigma}^2) p(\tilde{\boldsymbol{x}}) p(\boldsymbol{a}) p(\boldsymbol{b}_O) p(\mu_O) p(\sigma_O). \tag{A.8}$$

3 Application

In addition to a set of differentially corrected GPS points ($m = G$) and a large–scale (1:24000) USGS Digital Line Graph (DLG) ($m = D$) (U.S. Geological Survey 1999, 2004), each discussed in Section 2.2, we also use a Census 2000 TIGER/line file ($m = C$) (U.S. Census Bureau 2000a,b) and a StreetMap USA file ($m = S$) (Environmental Systems Research Institute 2003)—an "enhanced" version of the TIGER/line file. (TIGER is an acronym for Topologically Integrated Geographic Encoding and Referencing system.) Each source depicts a residential neighborhood in Durham, North Carolina, USA (Figure 2). Each has 44 points that represent the same 44 residential street intersections, 41 of which are depicted as intersections in the figure; along Watts Street ($l = 14$), three streets falling outside the boundaries of our maps create three points of intersection that are not depicted as such here, but are nonetheless point features that fall on line feature $l = 14$. All sources are in the same reference system, WGS 84 UTM Zone 17 meters. On–site inspection of the streets strongly suggests that they were intended to fall along straight lines. We have no instances of intersections involving three or more lines, hence no line parameters are determined in any of our line models (A.6), (A.7), or (A.8); see Section 2.4 and Appendix 4.

To better illustrate the effectiveness of the line model specification (A.6) compared to the points–only model (A.5) for inferring true position, we initially use only the Census 2000 TIGER/line file ($m = C$) and GPS points ($m = G$). To illustrate the effects of borrowing strength among slopes in line models, we use all four sources. In all illustrations, we use only $n_G = 16$ of the 44 available GPS points for model fitting, the remaining 28 GPS points being used for model checking; when using other sources, we always use all $n_m = N = 44$ observed points, $m \in \{C, D, S\}$. See Figure 2.

We use WinBUGS version 1.4.1 (Lunn et al. 2000) for all of the computations here unless noted otherwise.

3.1 Specifying prior distribution parameters

In Sections 2.2 and 2.3, we provided some discussion for specifying prior distributions. Here, we supply numerical values for prior hyperparameters that are particular to our application.

Source metadata suggests that, aside from GPS ($m = G$), the DLG ($m = D$) has the lowest positional uncertainty, while the Census 2000 TIGER/line file ($m = C$) and the StreetMap USA file ($m = S$) have similar uncertainties. (The StreetMap USA files are derived from the TIGER/line files.) As discussed in Section 2.2, the DLG has an *a priori* uncertainty corresponding to a total *a priori* positional error variance $\sigma^2_{t,\mathrm{D}} \approx 6^2\,\mathrm{m}^2$. We can derive a total *a priori* positional error variance for the TIGER/line file in a similar manner.

TIGER/line metadata states that each file

> *...at best meets the established National Map Accuracy Standards...where 1:100,000–scale maps are the source...*(U.S. Census Bureau 2000b).

This implies that TIGER/line source maps introduce a variance of about $24^2\,\mathrm{m}^2$, according to National Map Accuracy Standards; see see Section 2.2. Though we do not know the standards for producing TIGER/line files from such source maps, we assume, additionally, "0.003 inches standard error" due to the production process from the source, the same as for production of DLGs from their source maps (U.S. Bureau of the Budget 1947; U.S. Geological Survey 1999). This implies a variance of about $8^2\,\mathrm{m}^2$ at a map scale of 1:100000. Together, these two variances suggest a total *a priori* positional error variance, $\sigma^2_{t,\mathrm{C}} \approx 25^2\,\mathrm{m}^2$. Since StreetMap USA files are derived from TIGER/line files, we assume also $\sigma^2_{t,\mathrm{S}} \approx 25^2\,\mathrm{m}^2$. The processing software for GPS positions indicates a positional uncertainty corresponding to $\sigma^2_{t,\mathrm{G}} \approx 0.7^2\,\mathrm{m}^2$ (Trimble Navigation 1997a, b, 2003).

Thus, following the discussion at the end of Section 2.3, we specify independent scaled inverse–$\chi^2(\nu_m, s^2_m)$ priors for σ^2_{mq}, with each prior centered, respectively, at means given by the aforementioned values of $\sigma^2_{t,m}$, i.e., $\mathrm{E}(\sigma^2_{mq}) = \frac{\nu_m}{(\nu_m-2)} s^2_m \equiv \sigma^2_{t,m}$, $m \in \{C, D, G, S\}$. For $m \in \{C, D, S\}$, we assign degrees of freedom $\nu_m = 3$, giving weakly informative specifications in the sense that $\mathrm{Var}(\sigma^2_{mq})$ does not exist. Because GPS points are relatively precise compared to the remaining sources of positional data, we might argue, as a practical approximation, that $\sigma^2_{Gq} = \sigma^2_{t,G}$ with probability one, $q = 1, 2$. But, we allow for some uncertainty in σ^2_{Gq} and specify $\nu_G = 10$.

The map region $\mathcal{X}$ is contained approximately within a $1\,\mathrm{km}^2$ square area; see Figure 2. For some uniform prior specifications, it is convenient to work with the centered region contained approximately within $[-500, 500] \times [-500, 500]$, now using meters. The corresponding uniform

prior specifications for the original region are simply shifted versions of the centered specifications. For each stochastic true coordinate, we specify an independent uniform$(-625, 625)$, allowing for five maximum *a priori* map standard deviations ($\sigma_{t,C} \approx 25$ m) beyond the edges of the centered region. The coordinates of the "points," (x_{01}^l, x_{02}^l), $l = 1, \ldots, L$, in the point–slope specification of lines, are assigned the same independent uniform priors. Recall (Section 2.4) that intercepts follow from slopes and these points.

The remaining specifications are the same regardless of working with the centered region or the original. The affine shift parameters β_{m10} and β_{m20}, $m \in \{C, D, S\}$ are each assigned an independent uniform$(-125, 125)$, which, again, allows for plus or minus five times the maximum *a priori* map standard deviation ($\sigma_{t,C} \approx 25$ m). The remaining affine parameters for $m \in \{C, D, S\}$ are, independently, uniform$(0, 2)$ for β_{m11} and β_{m22}, and uniform$(-1, 1)$ for β_{m21} and β_{m12}; these bounds allow for plausible scaling or rotating of the affine transformation. Thus, we specify *a priori* accuracy in the sense that $\mathrm{E}(\boldsymbol{\beta}_m) \equiv E(\beta_{m10}, \beta_{m20}, \beta_{m11}, \beta_{m21}, \beta_{m12}, \beta_{m22})^T = (0, 0, 1, 0, 0, 1)^T$ (Section 2.3). As discussed in Section 2.3, we set $\boldsymbol{\beta}_G = (0, 0, 1, 0, 0, 1)^T$ with probability one.

For each of the slope scale parameters, σ_b of model (A.7) and σ_O of model (A.8), we specify a uniform$(0, 10)$, and we use parameter expansion (Gelman et al. 2003; Gelman 2006) to alleviate "sticky" sampling behavior due to σ_b and σ_O being near zero.

3.2 Specifying initial values

Here, we give some details on how to obtain initial sample values for all stochastic quantities. Our approach is fairly straightforward and intuitive, and is guided by the model specifications given in Sections 2.3 and 2.4.

Recall that our source maps do not depict more than two lines intersecting at any one point, so that all line parameters, a_l and b_l, $l = 1, \ldots, L$, are stochastic. Their initial values were obtained in the following manner. We chose the coordinates of the $n_G = 16$ GPS points, our *a priori* best set of points, to serve as dependent variables in two regressions. One regression used the $n_G = 16$ NS coordinates as dependent variable, the other regression used the $n_G = 16$ EW coordinates as dependent variable. We chose the next best set of points to be used as regressors. For illustration purposes, assume this set consists of observed DLG ($m = D$) points. (For some results, we use only the two sources $m = G$ and $m = C$, in which case, of course, the second best set would be those of source $m = C$.) In particular, we used the corresponding 16 coordinate pairs of source $m = D$ as regressors, with intercept. Thus, we performed a sort of *ad hoc* estimation of affine parameters, but these should *not* be used as intitial affine parameters as we explain below.

We used each of the above two coordinate regressions to predict $N = 44$ coordinate values corresponding to all 44 observed regressor points of $m = D$. Together, the two sets of predicted coordinates constitute 44 coordinate pairs that might be considered as estimates of true points if only for the fact that the estimates are not necessarily arranged on lines, though we know with which true line(s) each point is associated. To obtain initial values for the points and slopes in the point–slope form of line l, $l = 1, \ldots, 14$, we used the subset of the 44 predicted points that is associated with line l. Regressing the values of one coordinate of subset l on the values of the remaining coordinate in a simple linear regression gives slope and intercept estimates that were used as initial values for the point and slope in the point–slope form of line l. In models (A.7) and (A.8), which use a hierarchical specification of slopes, we used the

estimated slopes from the simple linear regressions to obtain initial values for the parent means μ_{NS}, μ_{EW}, and μ_O, and for the scale terms σ_b and σ_O.

In our application here, these true line parameter initial values determined lines whose intersections determined estimates of 41 of the $N = 44$ true points. The remaining three true points lie on a single line and, thus, only one coordinate of each of these points is determined, each remaining coordinate having a stochastic—recall, uniform—specification, hence requiring an initial value; see points 39, 40, and 42 on Watts Street in Figure 2. We simply predicted the three necessary initial coordinates from the aforementioned simple linear regression for $l = 14$.

Now, these $N = 44$ true point estimates—only three coordinates of which are used as initial values—can then be used to obtain initial values for the affine parameters $\boldsymbol{\beta}_m$. For β_{m10}, β_{m11}, and β_{m12}, we regressed the EW coordinates of the $n_m = 44$ observed points, $m \in \{C, D, S\}$, on both coordinates of the corresponding $N = 44$ estimated true points, as (A.2) suggests; for β_{m20}, β_{m21}, and β_{m22}, we used the $n_m = 44$ NS coordinates as response. Initial values for the data source variance components $\sigma^2_{m,q}$, $m \in \{C, G, D, S\}$, $q = 1, 2$, were obtained from the estimated regression standard errors.

For a solitary true point in line models (A.6–A.8)—we have no such points here—or for a point in the points–only model (A.5), an initial coordinate value may be equated to the average of corresponding observed coordinates or may be generated from a Gaussian distribution whose mean is equal to the average of corresponding observed coordinates and whose variance is equal to an average of $\sigma^2_{t,m}$ values.

Multiple sets of initial values were obtained by repeating the above sequence of steps, each set beginning with slopes and intercepts generated from the bivariate Gaussians with means given by the regression–estimated slope and intercept parameters, and variances given by the estimated variance–covariance matrices for the slope and intercept parameter estimators.

3.3 Results

We computed three parallel Markov chains from initial values obtained via the procedure outlined above. Generally speaking, convergence was achieved quickly, well within a few hundred iterations, and mixing was excellent. More formally, we assessed convergence using the potential scale reduction factor (Gelman and Rubin 1992) and its multivariate version (Brooks and Gelman 1997) as implemented in the CODA add–on package (Plummer et al. 2006) in the R Language and Environment for Statistical Computing (R Development Core Team 2006). Convergence was deemed to have occurred if each of the univariate factors and the multivariate version were less than 1.1. Each chain consisted of an initial burn–in set of 5000 iterations, which included any adapting phase, plus a subsequent set of 5000 iterations. We used the latter 5000 iterations of one chain for each model to compute all results reported here.

We use the deviance information criterion (DIC) (Spiegelhalter et al. 2002) and posterior predictive loss (Gelfand and Ghosh 1998) to compare models throughout. In particular, for the latter, we compute the posterior mean of

$$\frac{\sum_{mkq} \left(y^{rep}_{mq}(\boldsymbol{x}_{j_m(k)}) - y_{mq}(\boldsymbol{x}_{j_m(k)}) \right)^2}{2 \sum_m n_m},$$

where y^{rep}_{mq} is the predicted value of the observed coordinate y_{mq}, and the denominator puts the result on a per–coordinate basis for convenience of interpretation. We refer to this as

MSPE. We note that, while all DIC values reported here are associated with a positive *overall* effective number of parameters, all but the very first and very last reported DIC values below are associated with at least one factor in the likelihood that contributes a negative value for the effective number of parameters (Spiegelhalter et al. 2002).

Point model versus line model

Results for the points–only model (A.5) are illustrated in Figure (3) (DIC = 793.7, MSPE = 119.5 m^2). Recall that we used only sources $m = G$ and $m = C$ for this illustration. The figure shows the 16 relatively accurate GPS points used in the model fitting ($\times$) and the remaining 28 hold–out GPS points ($\boxtimes$). The figure also depicts approximate 95% credible ellipses for selected points separated by regular intervals along line segments connecting posterior mean estimates of the $N = 44$ true points. Each such coordinate pair along a line segment can be expressed as a convex combination of segment endpoints, giving a set of posterior draws for each point on a line segment, and the posterior mean and covariance matrix for each such point was estimated from these draws. Each ellipse is the 95% HPD credible region of an approximating bivariate Gaussian distribution using the estimated posterior mean and covariance matrix.

Analogous results for the comparable lines model (A.6) are illustrated in Figure (4) (DIC = 650.3, MSPE = 68.5 m^2). In this case, true points of intersection are estimated by the intersection of lines having posterior mean slope and mean intercept, the remaining three points fall along line feature $l = 14$ and are determined using the appropriate posterior mean coordinate in the equation for line $l = 14$. The differences in the precision of point–wise predictions of true position are remarkable compared to the results of the points–only model (A.5). Many lines' approximate 95% point–wise credible ellipses are comparatively imperceptible, the obvious exception being those for points along Broad Street, $l = 7$ in Figure 2, the second NS street from the left. This is the only street that was not directly associated with relatively precise GPS points; the relative imprecision of the Census 2000 TIGER/line file points is obvious. Loosely speaking, the NS coordinates of the intersection points along Broad Street are relatively constrained by their association with the EW streets, whose NS coordinates—as well as EW—are "tied–down" by the relative precision of GPS points, indicated by $\times$ in Figure 4. However, the association of Broad Street with EW streets appears to have comparatively little effect on the precision of the EW coordinates of Broad, hence little effect on the precision of the "EW location" of Broad. Still, we get a very noticeable improvement for this linear feature compared to the points–only model.

As a predictive check, we determined that all of the 28 hold–out GPS points fall within their respective 95% posterior predictive ellipses for the points–only model (A.5). For the lines model (A.6), 27 out of 28 GPS hold–out points fall in their respective credible regions for an observed coverage of 96.4%; GPS point 15 is not contained in its region. In each model, these predictive ellipses are centered at posterior predictive mean GPS locations, which, according to our model, are the same as the posterior mean true locations, but the predictive ellipses are slightly larger than those of the true points to account for the added variability of GPS about truth. Of course, we might expect one or two of these hold–out points to lie outside their 95% credible regions, and, as alluded to earlier, we do not believe that the points–only model realistically characterizes the variability of points along lines because it does not incorporate the linear process of linear features.

We note that an ellipse in Figures 3 or 4 characterizes the variability of an actual point feature only when associated with one of the $N = 44$ true points. Otherwise, the ellipses at

other locations along a line are a simple point–wise way to characterize variability of a line feature; credible regions for lines can be visualized as outlines that circumscribe collections of these point–wise ellipses. When using a line model, another way to express a credible region for a line feature as a whole would be to plot an ensemble or "bundle" of true line feature posterior draws such that the bundle consists of line features whose corresponding line's slope and intercept draws fall within the 95% HPD ellipse of an approximating bivariate Gaussian distribution; the outline of the resulting bundle would indicate a true line feature's approximate 95% credible region. In the current case, the ellipse approach and the bundle approach (not shown) give credible regions that are very similar.

Borrowing strength

The previous section presented conclusive results favoring the lines model (A.6) over the points–only model (A.5). Here, we look at the effect of borrowing strength among slopes in three different line models: the lines model (A.6), no borrowing of strength; the parallel lines model (A.7), two parent distributions for slopes; and the orthogonal lines model (A.8), one parent distribution for slopes of near–EW streets grouped together with negative inverse slopes of near–NS streets.

DIC for these three models is, respectively, 1634.3, 1632.2, and 1624.7, with corresponding MSPE values $60.26\,\mathrm{m}^2$, $60.87\,\mathrm{m}^2$, and $60.85\,\mathrm{m}^2$. (Recall that we use all four sources for these comparisons.) Thus, there is little practical difference among models according to MSPE, but DIC strongly suggests that streets exist in an orthogonal grid according to model (A.8). This conclusion is supported by the containment of zero in the 90% HPD credible interval of the posterior mean of the difference $\mu_{EW} - (-1/\mu_{NS})$ from model (A.7): $[-0.000854, 0.002945]$; the posterior histogram was unimodal and symmetric. (Of course, we might argue, instead, for a model with a single parameter—with standard Cauchy prior—for all slopes of EW streets and all negative inverse slopes of NS streets: DIC = 1654.1, MSPE = $61.0\,\mathrm{m}^2$; details omitted.)

The observed coverage for the GPS hold–out points by their 95% predictive credible ellipses are similar in each case: 27/28, 27/28, and 26/28, respectively; hold–out GPS point 15 was not covered for models (A.6) and (A.7), and, additionally, GPS point 8 was not covered for model (A.8). Thus, none of the three models indicate cause for concern regarding departure from nominal coverage.

We get corresponding improvements in marginal posterior precision of model parameters when comparing the models that borrow strength to the one that doesn't (Tables 1–4). This is apparent mostly for intercepts and slopes (Tables 1–2). And, we get improvements in posterior precision for true points, with posterior standard deviations of true EW coordinates averaging 0.6218, 0.4255, 0.4109 for models (A.6)-(A.8), respectively, and, for NS coordinates, 0.6156, 0.5284, 0.5357.

4 Discussion

We have presented a novel approach to combining multiple sources of positional information to infer true position. Moreover, we have shown how an accounting of feature linearity provides a remarkable improvement in precision for these inferences. The approach refines and extends the early notion of the epsilon–band characterization of

positional uncertainty in line segments (Perkal 1966; Chrisman 1982; Blakemore 1984; Veregin 1999) as well as the more recent, closely related refinements by various authors in the GIScience community (Caspary and Scheuring 1993; Leung and Yan 1998; Shi 1998; Shi and Liu 2000; Leung et al. 2004a). And, our approach is conceptually consistent with the measurement error (ME) model advocated by Goodchild (2004) and Leung, Ma and Goodchild (2004a; 2004b; 2004c; 2004d) in the context of measurement–based GIS (MBGIS) (Goodchild 1999), though our approach goes beyond this in its combining of multiple sources of information and the use of the line process model within a fully probabilistic framework.

Within the relatively large literature on spatial data analysis, to our knowledge, only the current approach and that of Barber et al. (2006) address the issue of positional error for the purpose of inferring true position. These approaches are important not only for directly inferring position, but, more generally, for incorporating positional uncertainty into functions of spatial coordinates, hence providing a model–based foundation of error propagation studies (Heuvelink, Burrough & Stein 1989; Heuvelink 1999). Moreover, the models may be relevant to some traditional spatial statistical analyses (Cressie 1993; Banerjee et al. 2004) of attributes identified by uncertain locations.

We anticipate a few concerns regarding our approach in this paper. Specifically, we attempt to discuss and allay concerns of identifiability and of comparability to the approach of Barber et al. (2006), hereafter referred to as BGS.

The informative prior specifications on the source model variance components, σ^2_{mq}, is important for identifying latent true positions. This, of course, may be viewed as a virtue of the Bayesian approach to incorporating genuine prior information. BGS address this specification of prior information in a manner similar to that presented here, though their models generally are different from ours. A feature unique to our approach is the line process specification, and, in going from the points–only model (A.5) to the lines model (A.6), for example, we go from 88 true coordinates to 28 line parameters plus three true coordinates. Furthermore, the considerations of model specification discussed in Section 2.4 and Appendix 4 help to ensure that the model specification is identifiable. Using the same data, BGS predict 88 true coordinates for each of their map models in a model averaging approach to predicting true position. Each of their map models use the corresponding observed map's coordinates as covariates in a large–scale affine mean model for each map, and this results in completely separate mean specifications across map models, model averaging being used only to combine predictions of true position. Here, we use the same true coordinates as covariates across maps, hence facilitating learning across maps.

The BGS approach seems to offer an advantage with regard to the implementation of spatial structure in the (residual) positional error vector, $\boldsymbol{\varepsilon}_m(\boldsymbol{x})$. If we were to specify a spatial process on each component of the positional error vector, then, evidently, $\boldsymbol{\Sigma}_m$ would be a function not only of the typical spatial correlation range or smoothness parameters, but also of the unknown true positions. As BGS remark, this would compound the practical problem of a typically weakly identifiable spatial correlation structure. As an approximation, we might instead use the observed source coordinates

as arguments of $\mathbf{\Sigma}_m$. Still, this approach would become computationally demanding for sources with a large number of points unless other approximations are made. A similar "big n" problem arises in the BGS approach. And, it seems that we would be forced to use the unknown true coordinates in order to predict one or more unobserved $\boldsymbol{y}_m(\boldsymbol{x}_j)$ points, that is, for predicting where a feature would exist on source m when the feature is not originally depicted on that source.

Though we presented some expressions in terms of a general structure for source model covariances, $\mathbf{\Sigma}_m$, we restricted our implementation to conditional independence of source coordinates within and across sources. Using the same data as we do here, BGS specify spatial structure within a different modeling framework, but their results suggest little to no spatial correlation in either the EW or NS error compentents for any of the sources they used and, furthermore, that the spatial processes within a source are not correlated. This may not always be the case, but these results support our independence assumption in the current case, and these results have important implications for the use of such models in practice: independence allows for relatively quick computations, which is critical to implementations involving more sources or sources with a large number of observed positions. Though not reported here, we did specify a spatial process for each coordinate of $\boldsymbol{\varepsilon}_m(\boldsymbol{x})$, $m = 1 \ldots, M$, using observed coordinates in $\mathbf{\Sigma}_m$ as an approximation to truth, but the marginal posteriors for the spatial ranges of correlation were highly sensitive to prior specifications, essentially mimicking the priors and giving poor sampling performance when attempting to specify vague priors. These observations are consistent with BGS results for these data.

Finally, in regard to BGS, we note that our points–only model (A.5) results are similar to BGS results, but our line process models (A.6–A.8) are not directly comparable to anything in the statistical or the GIScience literature.

Regarding MCMC sampling, we noticed poor behavior in some initial implementations of our models when near–NS linear features (see Figure 2) were parameterized as $x_2 = a_l + b_l x_1$; evidently, slopes "near" infinity were problematic. Though we maintain this parameterization throughout this article for clarity of presentation, we actually implemented the alternative parameterization, $x_1 = a_l^* + b_l^* x_2$, for near–NS features, where $a_l^* = -a_l/b_l$ and $b_l^* = 1/b_l$, $l = 6, \ldots, 14$, and as we present in Tables (1) and (2).

The results obtained here are encouraging with regard to future work on positional error modeling and inference for true position in GIS. In particular, map conflation remains an outstanding problem. According to current GIS practice, we manually identified points on different maps as corresponding to the same true point features. Hence the relatively small size of our data sets! However, current results now motivate investigation of automated, model–based map conflation for larger problems whereby information in the sources and metadata may be used to decide which points/lines correspond across maps. We might also attempt automation of the specification of linear features, i.e., automated model selection, again using available positional information to help decide which lines are straight and which are not, perhaps via DIC or some other measure. Ultimately, we anticipate development of a "data–model integration" approach to combining multiple sources of positional information with geospatial infor-

mation theory.

References

Arbia, G., Griffith, D. A., and Haining, R. P. (2003). "Spatial error propagation when computing linear combinations of spectral bands: The case of vegetation indices." *Environmental and Ecological Statistics*, 10(3): 375–396. 626

Atkinson, P. M. (1999). "Geographical information science: geostatistics and uncertainty." *Progress in Physical Geography*, 23(1): 134–142. 625

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, volume 101 of *Monographs on statistics and applied probability*. Boca Raton: Chapman & Hall/CRC. 625, 643

Barber, J. J. (2007). "Using rubber sheets to infer true map location." *CHANCE*, 1: 49–54. 626

Barber, J. J., Gelfand, A. E., and Silander, J. A., Jr. (2006). "Modeling map positional error to infer true feature location." *Canadian Journal of Statistics*, 34(4): 659–676. 626, 627, 631, 633, 643

Blakemore, M. (1984). "Generalisation and error in spatial data bases." *Cartographica*, 21: 131–139. 626, 643

Brooks, S. and Gelman, A. (1997). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7: 434–455. 640

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, volume 63 of *Mongraphs on Statistics and Applied Probability*. London: Chapman & Hall. 626

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Pacific Grove: Duxbury, 2 edition. 626

Caspary, W. and Scheuring, R. (1993). "Positional accuracy in spatial data bases." *Computational, Environmental, and Urban Systems*, 17: 103–110. 626, 643

Chrisman, N. R. (1982). "A theory of cartographic error and its measurement in digital databases." In Foreman, J. (ed.), *Proceedings of Auto–Carto*, volume 5, 159–168. Falls Church, VA: American Society of Photogrammetry and American Congress on Surveying and Mapping. 626, 643

Cressie, N. and Kornak, J. (2003). "Spatial statistics in the presence of location error with an application to remote sensing of the environment." *Statistical Science*, 18: 436–456. 626

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons, Inc., revised edition. 625, 643

Dowman, I. (1999). "Encoding and validating data from maps and images." In Longley et al. (1999b), 437–450. 631

Environmental Systems Research Institute (2003). *ESRI Data and Maps ArcGIS (v8.2) StreetMap USA Local Streets CD-ROM 8*. Redlands, CA: Environmental Systems Research Institute. 637

Fuller, W. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc. 626

Gabrosek, J. and Cressie, N. (2003). "The effect on attribute prediction of location uncertainty in spatial data." *Geographic Analysis*, 34: 262–285. 626

Gelfand, A. E. and Ghosh, S. K. (1998). "A minimum posterior predictive loss approach." *Biometrika*, 85(1): 1–11. 640

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–533. 637, 639

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC, 2nd edition. 629, 639

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7: 457–511. 640

Goodchild, M. and Gopal, S. (eds.) (1989). *The Accuracy of Spatial Data*. London: Taylor & Francis. 625

Goodchild, M. F. (1999). "Measurement–based GIS." In Shi, W., Goodchild, M., and Fisher, P. (eds.), *Proceedings of the International Symposium on Spatial Data Quality '99*, 1–9. Hong Kong: Hong Kong Polytechnic University. 627, 643

— (2004). "A general framework for error analysis in measurement-based GIS." *Journal of Geographical Systems*, 6(4): 323–324. 625, 627, 643

Guptill, S. C. and Morrison, J. L. (eds.) (1995). *Elements of Spatial Data Quality*. New York: Elsevier Science. 625

Heuvelink, G. B. (1999). "Propagation of error in spatial modelling with GIS." In Longley et al. (1999a), 207–217. 626

Heuvelink, G. B., Burrough, P. A., and Stein, A. (1989). "Propagation of errors in spatial modelling with GIS." *International Journal of Geographical Information Science*, 3(4): 303–322. 626

Leung, Y., Ma, J.-H., and Goodchild, M. F. (2004a). "A general framework for error analysis in measurement-based GIS Part 1: The basic measurement-error model and related concepts." *Journal of Geographical Systems*, 6(4): 325–354. 626, 627, 643

— (2004b). "A general framework for error analysis in measurement-based GIS Part 2: The algebra-based probability model for point-in-polygon analysis." *Journal of Geographical Systems*, 6(4): 355–379. 626, 627, 643

— (2004c). "A general framework for error analysis in measurement-based GIS Part 3: Error analysis in intersections and overlays." *Journal of Geographical Systems*, 6(4): 381–402. 626, 627, 643

— (2004d). "A general framework for error analysis in measurement-based GIS Part 4: Error analysis in length and area measurements." *Journal of Geographical Systems*, 6(4): 403–428. 626, 627, 643

Leung, Y. and Yan, J. (1998). "A locational error model for spatial features." *International Journal of Geographical Information Science*, 12(6): 607–620. 626, 643

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (eds.) (1999a). *Geographical Information Systems Management Issues and Applications*, volume 2. New York: John Wiley & Sons, Inc., 2 edition. 625, 646

— (1999b). *Geographical Information Systems Principles and Technical Issues*, volume 1. New York: John Wiley & Sons, Inc., 2 edition. 625, 646, 648

Lowell, K. and Jaton, A. (eds.) (1999). *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*. Chelsea, Michigan: Ann Arbor Press. 625

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS–A Bayesian modelling framework: Concepts, structure, and extensibility." *Statistics and Computing*, 10: 325–337. 637, 638

Mowrer, H. T. and Congalton, R. G. (eds.) (1999). *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*. Chelsea, Michigan: Ann Arbor Press. 625

Perkal, J. (1966). "On the length of empirical curves." In Nystuen, J. D. (ed.), *Discussion Paper Number 10*. Ann Arbor: Michigan Inter-University Community of Mathematical Geographers. 626, 643

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). *coda: Output analysis and diagnostics for MCMC*. R package version 0.10-5. 640

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 640

Raftery, A., Madigan, D., and Hoeting, J. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92: 179–191. 626

Shi, W. (1998). "A generic statistical approach for modelling error of geometric features in GIS." *International Journal of Geographical Information Science*, 12(2): 131–143. 626, 643

Shi, W., Fisher, P. F., and Goodchild, M. F. (eds.) (2002). *Spatial Data Quality*. New York: Taylor & Francis. 625

Shi, W. and Liu, W. (2000). "A stochastic process-based model for the positional error of line segments in GIS." *International Journal of Geographical Information Science*, 14(1): 51–66. 626, 643

Soler, T. and Marshall, J. (2002). "Rigorous transformation of variance-covariance matrices of GPS-derived coordinates and velocities." *GPS Solutions*, 6: 76–90. 630

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4): 583–639. 640, 641

Thapa, K. and Bossler, J. (1992). "Accuracy of spatial data used in geographic information systems." *Photogrammetric Engineering & Remote Sensing*, 58(6): 835–841. 625

Trimble Navigation (1997a). *TDC1 Asset Surveyor Operation Manual (31174-20)*. Sunnyvale, CA. 630

— (1997b). *TDC1 Asset Surveyor Software User Guide (31173-20)*. Sunnyvale, CA. 630

— (2003). *GPS Pathfinder Office Getting Started Guide (34231-30-ENG)*. Westminister, CO. 630

U.S. Bureau of the Budget (1947). "United States National Map Accuracy Standards." US Bureau of the Budget, http://rockyweb.cr.usgs.gov/nmpstds/nmas.html. Revision June 17, 1947. 630, 638

U.S. Census Bureau (2000a). "Redistricting Census 2000 TIGER/Line Files [machine-readable data files]." 637

— (2000b). "Redistricting Census 2000 TIGER/Line Files Technical Documentation." 637, 638

U.S. Geological Survey (1999). "Standards for Digital Line Graphs, National Mapping Program Technical Instructions 09/99." Department of the Interior, US Geological Survey National Mapping Division, http://edc2.usgs.gov/geodata/index.php. 629, 630, 637, 638

— (2004). "1:24,000 Scale Digital Line Graphs (DLG)." http://edc2.usgs.gov/geodata/index.php. January 20. 629, 637

Veregin, H. (1999). "Data quality parameters." In Longley et al. (1999b), 177–189. 625, 626, 643

Wolf, P. R. and Ghilani, C. D. (1997). *Adjustment computations: Statistics and least squares in surveying and GIS*. Wiley Series in Surveying and Boundary Control. New York: John Wiley & Sons Inc. 626

Zhang, J. and Goodchild, M. F. (2002). *Uncertainty in geographic information*. Research Monographs in Geographic Information Systems. London: Taylor & Francis. 625

Appendix

The following examples should help to clarify model specification for true points and true lines. The specification may be considered as a second stage in a hierarchical development whose first stage (Section 2.3), conditional on true points, is straightforward. The examples should also clarify that true points, conditional on true lines, are easy to specify. And, in most map applications, illustrated by our own application in Section 3 and by Case (i), below, all line parameters may be specified using priors as discussed in Section 2.4. However, Cases (ii) and (iii), below, illustrate that line modeling may be helped or hindered in more complicated line feature configurations, depending on which pairs of lines are chosen to determine true point intersections. Fortunately, these latter two cases illustrate the exception rather than the rule for most map applications.

Consider the three configurations of "streets" shown in Figure 5. In total, there are $L = 7$ or $L = 8$ true lines and $N = 13$ or $N = 15$ true points ($\times$), all but one true point ($j = 13$) occurring on at least two lines. There are three EW lines ($l = 1, 2, 3$) and four NS lines ($l = 4, 5, 6, 7$) that intersect in a grid–like fashion as we might expect to see in many street networks. Line $l = 8$ (middle and bottom configurations) is used to illustrate complications that may arise with points of intersection consisting of more than two lines.

Case (i): Line 8 absent (top configuration). The resulting configuration is very similar to that in our application (Figure 2), and a model specification is straightforward. Given true lines, point coordinates are easily specified—deterministically or stochastically—as previously discussed in Section 2.4; only one coordinate of one point (13) is not determined in this case. Lines are not determined by other lines (see below) and may be specified stochastically as in Section 2.4.

Case (ii): Line 8 passes through two or fewer 3–line (generally n–line $n \geq 3$) intersections (middle configuration intersections 1 and 5). In this case, a specification for true points, again, is straightforward, given true lines. But, now, we have more than one way to determine points 1 and 5. For example, we may use line pairs 1 and 4, and 2 and 5, respectively. That is, points 1 and 5 are functions of these lines' parameters as shown in Section 2.4. Then, because line 8 must be made to pass through these two points, it is clear that line 8, given lines 1, 4, 2, and 5, is determined by these four lines, though the dependence in WinBUGS can be made via the two true points; there is no need to derive explicitly the equation of line 8 in terms of the other lines' parameters. Remaining lines are not determined by other lines and may be specified stochastically as in Section 2.4.

Alternatively, we may choose to determine points 1 and 5 using line 8 with lines 1 and 2, respectively. In this case, given lines 1, 2, and 8, lines 4 and 5 are now partially determined by these lines because lines 4 and 5 pass through points 1 and 5. Thus, we

seem to have found a sensible specification for lines 4 and 5 via point–slope formulas; the "points" (1 and 5) are determined, as we said, by other lines, but the slopes of lines 4 and 5 are not determined by other lines, and these slopes may be specified stochastically as in Section 2.4. Remaining lines are not determined by other lines and may be specified stochastically as in Section 2.4.

Case (iii): Line 8 passes through more than two 3–line intersections (bottom configuration intersections 1, 5, and 9). As in Case (ii), point specification is easy, though we must choose which line pairs to use to determine points 1, 5, and 9. We may determine these points by lines 1 and 4, 2 and 5, and 3 and 6, respectively. But, then, similar to Case (ii), line 8 is determined by these lines, and we are led to consider three equations involving the parameters of line 8:

$$\begin{aligned} x_{12} &= a_8 + b_8 x_{11} \\ x_{52} &= a_8 + b_8 x_{51} \\ x_{92} &= a_8 + b_8 x_{51}, \end{aligned}$$

where $\boldsymbol{x}_1 = (x_{11}, x_{12})^T$, $\boldsymbol{x}_5 = (x_{51}, x_{52})^T$, and $\boldsymbol{x}_9 = (x_{91}, x_{92})^T$ have been determined by, i.e., are functions of, other lines. Thus, given these other lines, we have three equations and two unknowns so that a_8 and b_8 are over– determined. It is important to realize that it is not appropriate simply to toss out one of these equations since there is information to be had in the specification that line 8 passes through all three intersections; this is part of the process model (Section 1.2).

It turns out that this particular case is not so bad, however, as we can arrive at a clear, sensible specification by determining points using different line pairs. If we let points 1, 5, and 9 be determined instead by line 8 and, say, lines 4, 5, and 6, respectively, then there is no problem; assign a stochastic specification to these lines as in Section 2.4. Lines 1, 2, and 3 are now partially determined by other lines via points 1, 5, and 9 respectively, and these lines may be specified in the same point–slope manner as discussed in Case (ii).

In summary, points may be seen as easy to specify, given lines. But, in some cases, the way in which points are determined, i.e., the pair of lines chosen to determine a point, may help or hinder subsequent line specification. We suspect there to be a systematic approach for handling such cases. Notice, in Case (iii), for example, we were able to arrive at a sensible specification by reducing the determination of points 1, 5, and 9 from using six lines—1 and 4, 2 and 5, 3 and 6, respectively—to using four—4, 5, 6, and 8. If this one instance can be used to speculate, perhaps a general rule is to use as few lines as possible to determine points. But, this will remain speculation for us because we do not pursue the matter further since such cases in the map context are the exception rather than the rule.

Acknowledgments

The authors wish to thank an anonymous Associate Editor and Referee for their valuable comments.

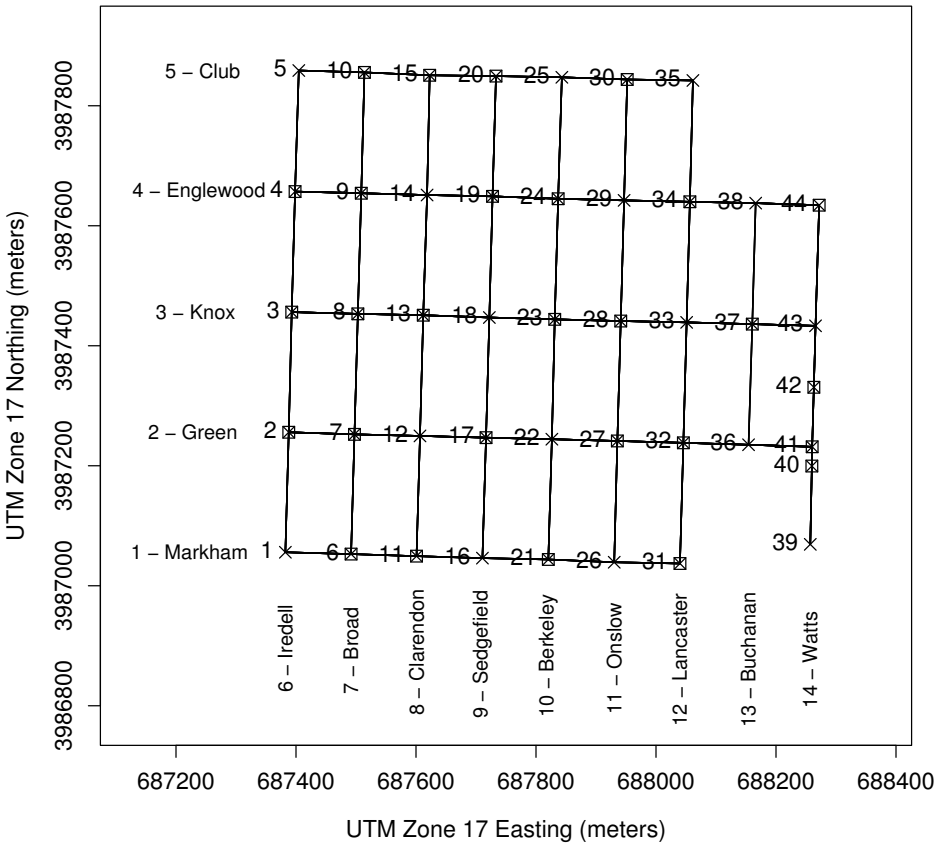

Figure 2: Map ($m = G$) of residential neighborhood streets in Durham, NC, USA, defined by 44 enumerated points. $\times$ indicates GPS points used in model fits, and $\boxtimes$ indicates GPS points used only for model checking. Corresponding 44 points of other sources (not shown) are used in all model fits. Street numbers correspond to line subscript, l.

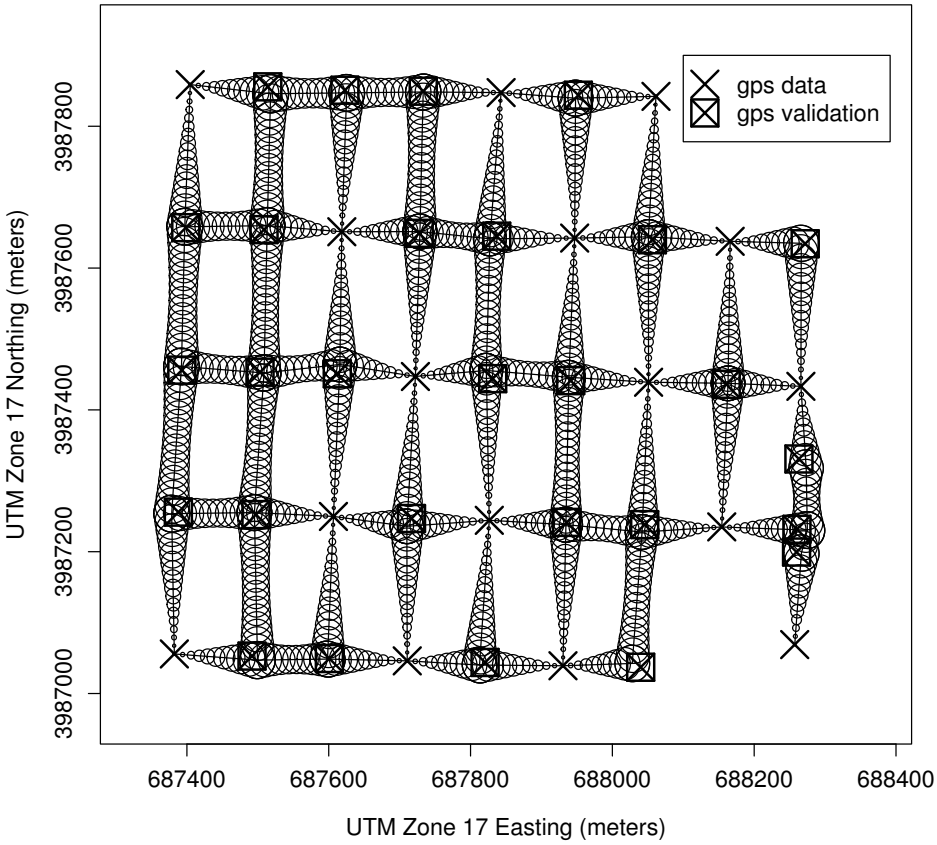

Figure 3: Points–only model (A.5) approximate 95% credible ellipses for selected points along line segments connecting true point posterior means. × indicates a GPS point used in the model fit, ⊠ indicates a GPS point only used for validation purposes. Compare to Figure 4.

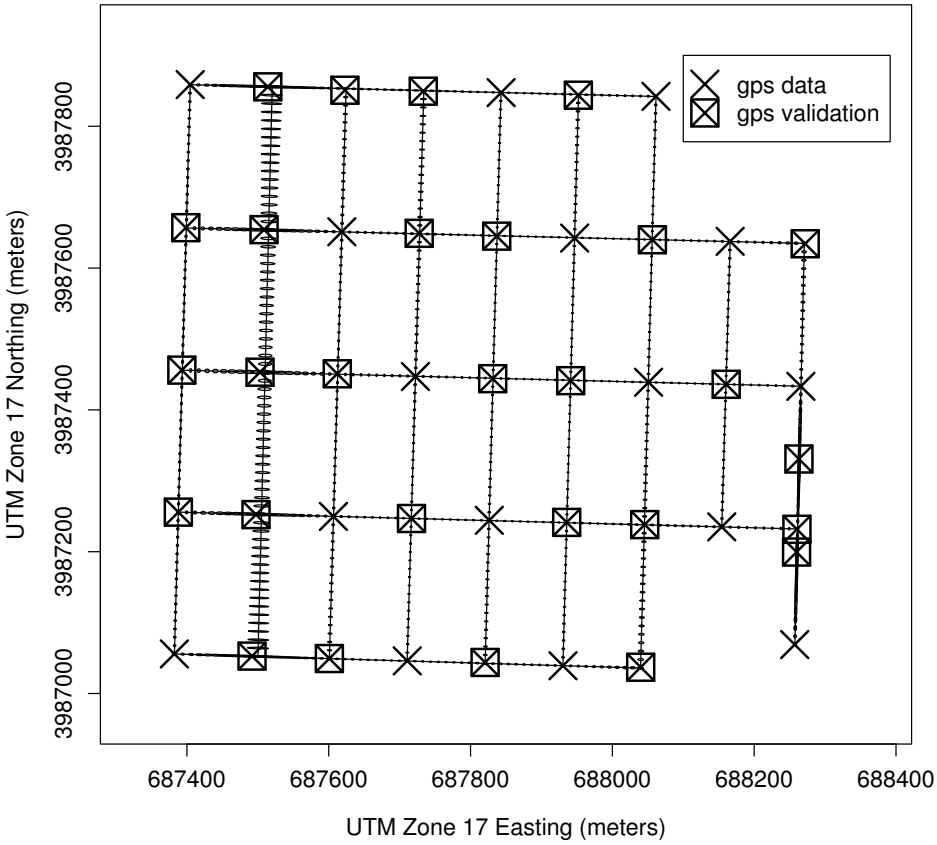

Figure 4: Lines model (A.6) approximate 95% credible ellipses for selected points along line segments connecting true point posterior means. $\times$ indicates a GPS point used in the model fit, $\boxtimes$ indicates a GPS point only used for validation purposes. Compare to Figure 3.

	mean	sd	2.5%	median	97.5%
Model (A.6)					
a_1	−382.5	0.5344	−383.5	−382.5	−381.4
a_2	−182.1	0.4022	−182.9	−182..0	−181.3
a_3	19.95	0.4355	19.14	19.92	20.84
a_4	220.6	0.3894	219.8	220.6	221.4
a_5	422.5	0.4333	421.7	422.5	423.4
a_6^*	−451.3	0.5518	−452.3	−451.3	−450.1
a_7^*	−337.0	1.904	−340.7	−337.0	−333.2
a_8^*	−232.2	0.5509	−233.0	−232.0	−230.8
a_9^*	−122.3	0.6493	−123.6	−122.3	−121.0
a_{10}^*	−12.87	0.5812	−14.14	−12.83	−11.84
a_{11}^*	96.64	0.5418	95.46	96.67	97.64
a_{12}^*	206.1	0.7408	204.5	206.1	207.5
a_{13}^*	315.6	0.5653	314.4	315.7	316.6
a_{14}^*	422.2	0.7731	420.8	422.1	423.9
Model (A.7)					
a_1	−382.5	0.4091	−383.3	−382.5	−381.7
a_2	−182.0	0.3796	−182.8	−182.0	−181.3
a_3	19.83	0.3914	19.12	19.81	20.65
a_4	220.6	0.3653	219.9	220.6	221.3
a_5	422.4	0.3954	421.6	422.4	423.1
a_6^*	−451.3	0.4556	−452.2	−451.3	−450.4
a_7^*	−337.2	1.833	−340.7	−337.2	−333.4
a_8^*	−232.0	0.4389	−232.8	−232.0	−231.1
a_9^*	−122.5	0.4741	−123.4	−122.5	−121.6
a_{10}^*	−12.76	0.4682	−13.76	−12.73	−11.91
a_{11}^*	96.73	0.4479	95.79	96.74	97.60
a_{12}^*	206.2	0.4917	205.2	206.2	207.1
a_{13}^*	315.7	0.4524	314.8	315.7	316.6
a_{14}^*	422.7	0.5217	421.7	422.6	423.8
Model (A.8)					
a_1	−382.6	0.4124	−383.3	−382.6	−381.7
a_2	−182.0	0.3907	−182.9	−182.0	−181.3
a_3	19.88	0.3806	19.15	19.88	20.66
a_4	220.6	0.3636	219.9	220.6	221.3
a_5	422.3	0.3999	421.5	422.3	423.1
a_6^*	−451.3	0.4481	−452.2	−451.3	−450.4
a_7^*	−337.2	1.812	−340.8	−337.2	−333.6
a_8^*	−232.0	0.4471	−232.9	−232.0	−231.1
a_9^*	−122.5	0.4618	−123.4	−122.6	−121.6
a_{10}^*	−12.74	0.456	−13.66	−12.74	−11.86
a_{11}^*	96.72	0.4458	95.79	96.74	97.58
a_{12}^*	206.3	0.4759	205.3	206.3	207.2
a_{13}^*	315.7	0.4553	314.8	315.8	316.6
a_{14}^*	422.6	0.4967	421.7	422.6	423.6

Table 1: Line intercepts posterior summary. Intercepts for near–NS features are reported using the line parameterization $x_1 = a_l^* + b_l^* x_2$, where $a_l^* = -a_l/b_l$ and $b_l^* = 1/b_l$, $l = 6, \ldots, 14$. Add 3987425 m to a_l and 687844 m to a_l^* to match the coordinate system in Figures 2–4. Subscripts refer to linear feature (street) number in Figure 2.

	mean	sd	2.5%	median	97.5%
Model (A.6)					
b_1	−0.02751	0.001864	−0.03089	−0.02759	−0.02353
b_2	−0.02681	0.001454	−0.02972	−0.02682	−0.02397
b_3	−0.02783	0.001473	−0.03082	−0.02778	−0.02508
b_4	−0.02735	0.001664	−0.03076	−0.02726	−0.02422
b_5	−0.02463	0.001469	−0.02740	−0.02465	−0.02155
b_6^*	0.02764	0.001367	0.02479	0.02767	0.03032
b_7^*	0.02710	0.005930	0.01602	0.02710	0.03907
b_8^*	0.02999	0.002495	0.02534	0.02992	0.03532
b_9^*	0.02930	0.002453	0.02464	0.02919	0.03455
b_{10}^*	0.02811	0.001785	0.02478	0.02803	0.03192
b_{11}^*	0.02753	0.001691	0.02425	0.02750	0.03104
b_{12}^*	0.02788	0.002470	0.02336	0.02778	0.03309
b_{13}^*	0.02945	0.002662	0.02391	0.02949	0.03482
b_{14}^*	0.02369	0.002852	0.01792	0.02374	0.02936
Model (A.7)					
b_1	−0.02707	0.001000	−0.02915	−0.02703	−0.02523
b_2	−0.02682	0.0009714	−0.02881	−0.02681	−0.02500
b_3	−0.02706	0.0009748	−0.02917	−0.02702	−0.02528
b_4	−0.02693	0.001033	−0.02907	−0.02692	−0.02497
b_5	−0.02639	0.001046	−0.02827	−0.02644	−0.02410
b_6^*	0.02788	0.0007946	0.02636	0.02787	0.02946
b_7^*	0.02786	0.001113	0.02576	0.02781	0.03039
b_8^*	0.02809	0.0009900	0.0264	0.02801	0.03040
b_9^*	0.02791	0.0009612	0.02604	0.02791	0.02993
b_{10}^*	0.02781	0.0008122	0.02621	0.02779	0.02950
b_{11}^*	0.02772	0.0008399	0.02599	0.02775	0.02939
b_{12}^*	0.02782	0.0009211	0.02604	0.02781	0.02974
b_{13}^*	0.02796	0.0009907	0.02618	0.02791	0.03006
b_{14}^*	0.02747	0.001037	0.02485	0.02756	0.02924
μ_{EW}	−0.02682	0.0008936	−0.02857	−0.02681	−0.02510
μ_{NS}^*	0.02782	0.0007102	0.02647	0.02780	0.02929
Model (A.8)					
b_1	−0.02766	0.0008312	−0.02944	−0.02762	−0.02608
b_2	−0.02741	0.0007141	−0.02879	−0.02744	−0.02592
b_3	−0.02751	0.0007627	−0.02917	−0.02748	−0.02608
b_4	−0.02742	0.0008235	−0.02913	−0.02742	−0.02573
b_5	−0.02688	0.0009119	−0.02834	−0.02702	−0.02475
b_6^*	0.02752	0.0006809	0.02618	0.02751	0.02893
b_7^*	0.02744	0.0009340	0.02539	0.02745	0.02938
b_8^*	0.02790	0.0009650	0.02636	0.02777	0.03024
b_9^*	0.02775	0.0008415	0.02623	0.02769	0.02969
b_{10}^*	0.02760	0.0007816	0.02614	0.02754	0.02933
b_{11}^*	0.02743	0.0007423	0.02590	0.02743	0.02887
b_{12}^*	0.02738	0.0008374	0.02553	0.02741	0.02907
b_{13}^*	0.02759	0.0008847	0.02595	0.02753	0.02953
b_{14}^*	0.02710	0.0009553	0.02475	0.02722	0.02871
μ_O	−0.02745	0.0005367	−0.02849	−0.02746	−0.02640

Table 2: Line slopes posterior summary. Slopes for near–NS features are reported using the line parameterization $x_1 = a_l^* + b_l^* x_2$, where $a_l^* = -a_l/b_l$ and $b_l^* = 1/b_l$, $l = 6, \ldots, 14$. Similarly, $\mu_{NS}^* = 1/\mu_{NS}$. Numerical subscripts refer to linear feature (street) number in Figure 2.

	mean	sd	2.5%	median	97.5%
Model (A.6)					
β_{210}	26.20	1.084	24.11	26.20	28.32
β_{211}	1.011	0.003943	1.003	1.011	1.019
β_{212}	0.0005256	0.004040	−0.007444	0.0005643	0.008524
β_{220}	−7.182	1.141	−9.416	−7.184	−4.918
β_{221}	0.008601	0.003959	0.0008044	0.008613	0.01642
β_{222}	1.016	0.004178	1.008	1.016	1.024
β_{310}	1.673	0.7706	0.1849	1.667	3.227
β_{311}	1.002	0.002840	0.9965	1.002	1.008
β_{312}	0.001247	0.002914	−0.004570	0.001196	0.007119
β_{320}	1.300	0.3360	0.6354	1.300	1.949
β_{321}	−0.001674	0.001263	−0.004226	−0.001646	0.0008307
β_{322}	0.9966	0.001191	0.9942	0.9966	0.9990
β_{410}	27.12	1.254	24.68	27.12	29.60
β_{411}	1.006	0.004377	0.9970	1.006	1.014
β_{412}	−0.0002720	0.004463	−0.009036	−0.0002844	0.008457
β_{420}	−6.056	1.070	−8.220	−6.044	−3.938
β_{421}	0.002162	0.003789	−0.005165	0.002202	0.009648
β_{422}	1.011	0.003945	1.003	1.011	1.018
Model (A.7)					
β_{210}	26.22	1.100	24.09	26.22	28.39
β_{211}	1.011	0.003857	1.003	1.011	1.018
β_{212}	0.0006714	0.003922	−0.006846	0.0006488	0.008327
β_{220}	−7.127	1.127	−9.356	−7.106	−4.915
β_{221}	0.008565	0.003981	0.0005690	0.008564	0.01637
β_{222}	1.016	0.004157	1.008	1.016	1.024
β_{310}	1.687	0.7530	0.1925	1.692	3.147
β_{311}	1.002	0.002701	0.9964	1.002	1.007
β_{312}	0.001355	0.002660	−0.003890	0.001407	0.006536
β_{320}	1.328	0.3239	0.6999	1.331	1.969
β_{321}	−0.001797	0.001243	−0.004212	−0.001829	0.0007118
β_{322}	0.9966	0.001164	0.9943	0.9966	0.9988
β_{410}	27.10	1.217	24.68	27.09	29.53
β_{411}	1.006	0.004376	0.9970	1.006	1.014
β_{412}	−0.0001070	0.004491	−0.009018	−0.00001055	0.008487
β_{420}	−6.033	1.045	−8.075	−6.033	−3.994
β_{421}	0.001967	0.003812	−0.005652	0.001956	0.009418
β_{422}	1.011	0.003972	1.003	1.011	1.019
Model (A.8)					
β_{210}	26.20	1.070	24.06	26.20	28.30
β_{211}	1.011	0.003873	1.003	1.011	1.018
β_{212}	0.0009734	0.003938	−0.006686	0.001070	0.008591
β_{220}	−7.124	1.105	−9.274	−7.133	−4.880
β_{221}	0.009021	0.003954	0.001079	0.009022	0.01652
β_{222}	1.016	0.004108	1.008	1.016	1.024
β_{310}	1.659	0.7556	0.1436	1.661	3.168
β_{311}	1.002	0.002636	0.9966	1.002	1.007
β_{312}	0.001625	0.002645	−0.003645	0.001648	0.006876
β_{320}	1.326	0.3249	0.6746	1.331	1.946
β_{321}	−0.001342	0.001099	−0.003472	−0.00135	0.0008238
β_{322}	0.9965	0.001147	0.9943	0.9965	0.9988
β_{410}	27.13	1.216	24.74	27.12	29.50
β_{411}	1.006	0.004355	0.9972	1.006	1.014
β_{412}	0.0001880	0.004424	−0.008442	0.0001630	0.008920
β_{420}	−6.028	1.070	−8.123	−6.024	−3.930
β_{421}	0.002376	0.003720	−0.005012	0.002393	0.009572
β_{422}	1.011	0.003933	1.003	1.011	1.018

Table 3: Affine parameters posterior summary.

	mean	sd	2.5%	median	97.5%
Model (A.6)					
σ^2_{11}	0.7568	0.2216	0.4556	0.7134	1.307
σ^2_{12}	0.7635	0.1808	0.4817	0.7389	1.178
σ^2_{21}	6.889	0.7802	5.526	6.809	8.599
σ^2_{22}	7.251	0.8205	5.882	7.175	9.108
σ^2_{31}	4.692	0.5434	3.793	4.638	5.931
σ^2_{32}	1.714	0.2173	1.348	1.694	2.185
σ^2_{41}	7.870	0.8818	6.372	7.789	9.817
σ^2_{42}	6.801	0.7665	5.510	6.728	8.495
Model (A.7)					
σ^2_{11}	0.6385	0.1333	0.4417	0.6171	0.9611
σ^2_{12}	0.7136	0.1385	0.497	0.6959	1.040
σ^2_{21}	6.966	0.7801	5.626	6.890	8.690
σ^2_{22}	7.276	0.8125	5.889	7.214	9.106
σ^2_{31}	4.636	0.5223	3.744	4.591	5.789
σ^2_{32}	1.716	0.206	1.364	1.697	2.184
σ^2_{41}	7.860	0.8648	6.379	7.790	9.725
σ^2_{42}	6.842	0.7734	5.554	6.774	8.523
σ^2_b	0.0006736	0.0005306	0.00002733	0.0005566	0.001972
Model (A.8)					
σ^2_{11}	0.6278	0.1295	0.4400	0.6074	0.9389
σ^2_{12}	0.7187	0.1388	0.4988	0.7026	1.037
σ^2_{21}	6.967	0.7892	5.666	6.887	8.719
σ^2_{22}	7.235	0.7996	5.885	7.158	8.957
σ^2_{31}	4.618	0.5205	3.747	4.567	5.785
σ^2_{32}	1.710	0.2020	1.366	1.694	2.151
σ^2_{41}	7.882	0.8829	6.390	7.801	9.826
σ^2_{42}	6.848	0.7607	5.527	6.777	8.562
σ^2_O	0.0006664	0.0005277	0.00002327	0.0005601	0.001968

Table 4: Standard deviation parameters posterior summary.

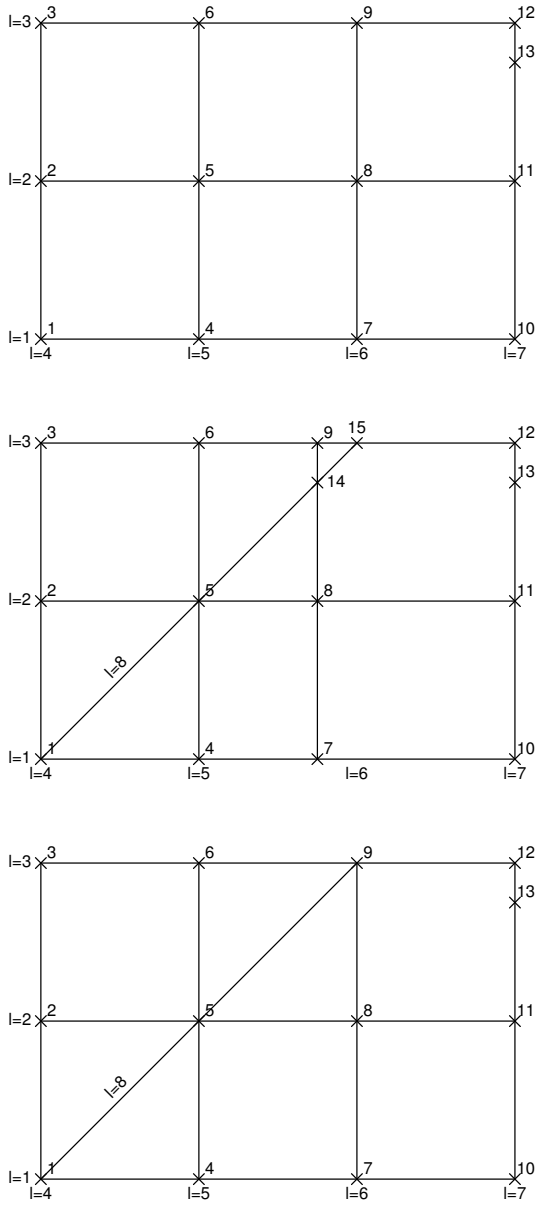

Figure 5: Three configurations used to illustrate model specification for true points and lines.