Modeling the probability distribution of positional errors incurred by residential address geocoding

Dale L. Zimmerman[1], Xiangming Fang[2], Soumya Mazumdar[3], and Gerard Rushton[4]

November 16, 2006

[1]Dale L. Zimmerman is Professor, Department of Statistics and Actuarial Science and Department of Biostatistics, University of Iowa, Iowa City, IA 52242 (E-mail: dzimmer@stat.uiowa.edu; Phone: 319-335-0818; Fax: 319-335-3017), and Affiliate of the Center for Health Policy and Research, College of Public Health, University of Iowa. Please address all correspondence to him.

[2]Xiangming Fang is a Ph.D. student in the Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242 (E-mail: xiangming-fang@uiowa.edu).

[3]Soumya Mazumdar is a Ph.D. student in the Department of Geography, University of Iowa, Iowa City, IA 52242 (E-mail: soumyamazumdar@yahoo.com).

[4]Gerard Rushton is Professor, Department of Geography, University of Iowa, Iowa City, IA 52242 (E-mail: gerard-rushton@uiowa.edu).

Abstract

Background: The assignment of a point-level geocode to subjects' residences is an important data assimilation component of many geographic public health studies. Often, these assignments are made by a method known as automated geocoding, which attempts to match each subject's address to an address-ranged street segment georeferenced within a streetline database and then interpolate the position of the address along that segment. Unfortunately, this process results in positional errors. Our study sought to model the probability distribution of positional errors associated with automated geocoding (both 100%-matched and 60%-matched) and E911 geocoding.

Results: Positional errors were determined for 1423 rural addresses in Carroll County, Iowa as the vector difference between each 100%-matched automated geocode and its true location as determined by orthophoto and parcel information. Errors were also determined for 1449 60%-matched geocodes and 2354 E911 geocodes. Huge (> 15 km) outliers occurred among the 60%-matched geocoding errors; outliers occurred for the other two types of geocodes also but were much smaller. E911 geocoding was more accurate (median error length = 44 m) than 100%-matched automated geocoding (median error length = 168 m). The empirical distributions of errors for all three geocodes exhibited a distinctive Greek-cross shape and had many other interesting features that were not capable of being fitted adequately by a single bivariate normal or t distribution. However, mixtures of t distributions with two or three components were able to fit errors corresponding to the 100%-matched and E911 geocodes very well.

Conclusion: Mixtures of bivariate t distributions with few components appear to be flexible enough to fit many positional error datasets associated with geocoding, yet parsimonious enough to be feasible for nascent applications of measurement-error methodology to spatial epidemiology.

Key words: Geocode, Location uncertainty, Mixture distributions, Positional accuracy, Spatial epidemiology.

Background

It is becoming increasingly common in public health studies to use the spatial locations of study participants in statistical analyses, for example to test for geographic clustering of disease or to estimate relationships between environmental exposures and disease. Indeed, statistical methods for spatial epidemiology are developing rapidly, and the growing list of book-length treatments of the subject include [1-4]. In order to utilize subjects' locations in a spatial analysis, it is necessary, of course, to define and ascertain these locations. Historically, the spatial location of a person has been defined as the person's place of residence; however, recognition of human mobility and the fact that many causative exposures occur outside the home have generated recent attempts to expand this definition to daily activity spaces and such constructs as time geography and pathogenic paths; for a brief review see [5]. Nevertheless, place of residence currently remains the typical representation of each subject's location in public health studies.

The spatial coordinates of a place of residence are usually not measured directly; rather, the residential address is given a location reference, known as a geocode. The geocode may be defined as the latitude and longitude coordinates or a point in some other coordinate system, or as a statistical tabulation area such as a U.S. Census tract, block group, or block; here, unless noted otherwise, we use the point rather than areal definition. Several distinct methods for geocoding exist, including visiting the residence with global positioning system (GPS) transmitters, identifying the residence on orthophoto maps based on aerial imagery, and matching the address to a digital street map. The latter can be done in batch mode for large numbers of addresses and when done this way is often called "automated geocoding." Recently, a new method of automated geocoding has been developed that matches an address to parcel descriptions of legal property boundaries developed by assessors, but this method has not yet been widely adopted. The U.S. Census Bureau is developing such a parcel-level geocode for all U.S. addresses, but the public does not and will not have access to these geocodes. Accordingly, automated geocoding here will refer to the widely used practice of using a geographic information system (GIS) to match an address to a street name and address range in a digitized street reference map and then estimate, via interpolation, where the address is located between the two points that define the limits of the address range.

Automated geocoding is cheaper, more convenient, and hence much more common than non-automated methods, but considerably less accurate. Several investigations of the accuracy of automated geocoding have recently been published. Some of these have measured accuracy by the proportion of addresses for which the geocode belongs to a correct statistical tabulation area; for example, Yang et al. [6] and Kravets and Hadden [7] found that only 70% to 90% of their geocoded addresses were assigned to the correct census block. Other investigations have measured accuracy by the Euclidean distance between the point location ascertained by automated geocoding and the corresponding "true" location as determined by a much more intensive and accurate method (e.g. GPS transmitters or aerial imagery) [8-13]. These latter studies have shown that positional errors of several hundred meters are incurred regularly by automated geocoding, and that even larger errors are not uncommon in rural areas. In one of the most thorough studies of automated geocoding errors published to date, Cayo and Talbot [14] found that 10% of a sample of rural addresses in a four-county upstate New York study area geocoded with errors of more than 1.5 km, and 5% geocoded with errors exceeding 2.8 km.

An alternative method of geocoding that may have promise for public health research is E911 geocoding. E911 geocodes are usually obtained under the auspices of local governments for the specific purpose of dispatching emergency vehicles to the correct location in response to a 9-1-1 telephone call requesting assistance. The particular methods used to obtain the geocodes vary, but they generally are more resource-intensive than mere automated geocoding due to the life-and-death issues at stake. For example, some counties have used parcel address-matching, while others have hired commercial firms that claim to take a GPS measurement at or near each residence. Every year, more counties in the U.S. develop E911 geocodes, so it is possible that in the not-too-distant future, many health researchers will be able to use these geocodes in lieu of performing automated geocoding. Investigations of the accuracy of E911 geocodes have not yet appeared in the scientific literature, though commercial firms offering E911 geocoding services tout them, unsurprisingly, as much more accurate than geocodes obtained via automated geocoding.

Whatever process is used to obtain geocodes of residences, the positional errors incurred by that process introduce location uncertainties that may adversely affect spatial analytic methods. Specific effects of positional errors on spatial statistical analyses include inflation of standard errors of parameter estimates and a reduction in power to detect such spatial features as clusters and trends [15-17]. Even relatively small positional errors can have a discernible impact on local statistics for detecting clustering or "hot spots" [18]. It is important, therefore, for researchers to quantify these effects on their analyses, which in turn requires them to have, or gain, some understanding of the probability distribution of the positional errors. In fact, the adoption of an adequate model for the distribution of positional errors is essential for successful implementation of existing measurement-error model methods for spatial data analysis; see, e.g., [19-22]. Knowledge of the error distribution also facilitates the use of multiple imputation methods for adjusting spatial statistical analyses for positional errors. These methods proceed by imputing (simulating) locations with error from the distribution of an observed location given its corresponding true location, whereupon inferences for the spatially-varying health outcome of interest can be made using the model for that outcome given the true locations, but with each true location replaced by multiple imputed realizations. Finally, gaining an understanding of typical geocoding error distributions allows for the simulation of realistic positional errors for power studies of various tests for clusters, spatial trends, and other important spatial patterns and features.

The main purpose of this article is to formulate and fit useful models for the probability distribution of positional errors incurred by geocoding residential addresses. In particular, we will formulate models that are sufficiently flexible to allow for the representation of features observed in empirical distributions of positional errors derived from a dataset of rural Iowa addresses, yet sufficiently simple that the aforementioned measurement-error and multiple imputation methodologies could be successfully implemented using these models. Positional errors corresponding to both automated geocoding and E911 geocoding will be considered. Upon formulating a suitable model or class of models for the errors, we will demonstrate how to fit those models to the data. Although the specific features seen in the distributions of positional errors from this predominantly rural Iowa county will not occur in all datasets, nor even in all error datasets derived from rural addresses, we believe that the methods we use to formulate and fit the models are generalizable to a great many datasets of positional errors incurred by geocoding.

In seeking useful models for a distribution of positional errors, one might first consider a bivariate normal distribution or a uniform distribution on a "standard" two-dimensional region (e.g. a circle or square). Indeed, normal and uniform distributions have been used previously to study the effects of location errors on spatial analyses in general, and on spatial prediction (kriging) and cluster detection in particular [23, 16, 19, 20]. However, to the authors' knowledge no empirical evidence has ever been presented to demonstrate that these distributions adequately represent the probability distributions of positional errors corresponding to geocoded residential addresses. In fact, these relatively simple distributions will not be appropriate if, for instance, extremely large positional errors (outliers) occur more often than would be expected for a bivariate normal or uniform distribution, or if errors tend to cluster along more than one axial direction. We will show that outliers and "multi-axial clustering" both occur for the positional errors in our geocoded data, and thus simple normal or uniform distributions will not suffice. As alternatives, we will propose the use of finite mixture distributions [24-26]. In a finite mixture distribution, each error can be regarded as having arisen from a population G which is a mixture of a finite number, say g, of subpopulations $G_1, \dots, G_g$ in some proportions $p_1, \dots, p_g$, respectively, where $\sum_{i=1}^{g} p_i = 1$ and $p_i \geq 0$ $(i = 1, \dots, g)$. The probability density function (pdf) of an arbitrary positional error, $\mathbf{x}$, can then be represented in the finite mixture form,

$$f(\mathbf{x};\boldsymbol{\phi}) = \sum_{i=1}^{g} p_i f_i(\mathbf{x};\boldsymbol{\theta}) \tag{1}$$

where $f_i(\mathbf{x};\boldsymbol{\theta})$ is the pdf corresponding to G_i; $\boldsymbol{\theta}$ denotes the vector of all unknown parameters associated with the parametric forms adopted for these g component pdfs; and $\boldsymbol{\phi} = (\mathbf{p}', \boldsymbol{\theta}')'$ where $\mathbf{p}' = (p_1, \ldots, p_g)$. Furthermore, we focus on mixtures of bivariate normal and t distributions, which are the most commonly used mixture models for bivariate observations and are well-suited for observations contaminated by outliers and exhibiting multi-axial clustering. The t mixtures are more robust than normal mixtures to contamination by outliers, hence they generally yield more parsimonious models than normal mixtures for data with outliers.

Methods

Data

The address data upon which this investigation is based consist of all 2516 rural residential addresses in Carroll County, Iowa, USA, current as of 31 December 2005, which we obtained in conjunction with a comprehensive study of rural health in Iowa by the Iowa Department of Public Health and other researchers at the University of Iowa. A major objective of the study was to investigate the possible existence of associations between various health outcomes and exposure to environmental contaminants produced by concentrated animal feeding operations. Hence the focus on rural addresses, which were defined as all residential addresses that lie outside incorporated township boundaries.

For this investigation, an attempt was made to obtain a geocode of each rural address using an automated method, an E911 method, and an orthophoto method, as described in the next subsection. The orthophoto method is very accurate, hence the geocodes produced by this method were taken as the truth. For each of the other two methods, the positional error corresponding to a given address was determined as the vector difference of the address's geocode obtained by the method and that address's orthophoto-derived geocode. For various reasons — most frequently the inability to determine which of several buildings in the photograph was the residence — a completely reliable orthophoto-derived geocode could not be ascertained for 162 of the addresses, so our analysis of positional errors is based on the remaining 2354 addresses.

Geocoding

Every address was assigned an automated geocode, an E911 geocode, and an orthophoto-derived geocode. We describe these geocoding processes in some detail below, as the details will be seen to have a bearing on the nature of the positional errors incurred by each process.

Automated geocoding. Addresses were matched to the U.S. Census Bureau's TIGER street centerline file for Carroll County using the GIS package ArcGIS 9.1 [27]. The geocoding process begins with automated parsing and standardization of the address list. Parsing is the process of breaking the address records up into distinct address component fields such as house number and street name, while standardization modifies these components, if necessary, so that they adhere to a common United States Postal Service standard [28]. Next, an address-ranged street segment in the TIGER file is probabilistically matched to each address on the basis of a "match score," which measures how closely each candidate address-ranged street segment in the TIGER file matches the address. Each field in the candidate segment is compared with the corresponding field of the address record being matched. The match score is a weighted composite score over all fields, scaled to lie between 0 to 100. For this analysis the minimum match score was set at either 100% (perfect matching) or 60%. Finally, the geocode is calculated by linearly interpolating the address number to a point on the matched street segment between the two points that define the limits of that segment's address range. No offset from the street centerline was used in this calculation.

E911 geocoding. For emergency services dispatch purposes, geocodes of all addresses in Carroll County are continually updated and maintained by the county government so that a 911 telephone caller within the county requesting assistance may be quickly and unambiguously located. The most suitable geocode for this purpose in rural areas was deemed by county officials to be the coordinates of the location where emergency service personnel would leave the public road and enter the private road leading to the property from which the call was made. We obtained these geocodes directly from the GIS coordinator of Carroll County, who was not able to say exactly how the contractor employed by Carroll County obtained them.

Orthophoto-based geocoding. Using visual identification, the third author enhanced the E911 geocode for each address to a location centered on the residence related to the address. This task was accomplished with the aid of 24 inch/pixel grayscale orthophotos of the study area we obtained from the Carroll County GIS Administrator and color infrared orthophotos (with the same resolution) obtained from the Iowa Natural Resources Geographic Information Systems Library (http://www.igsb.uiowa.edu/nrgislibx/). A GIS data layer indicating the parcel to which a particular property belonged (and which is used by the county assessor's office for tax assessment) was overlaid upon the orthophoto and E911 address layers to confirm that each geocode was assigned to the correct address.

Estimation of parameters

For each of the three sets of positional errors just described, we obtained likelihood-based estimates of the parameters of normal mixtures and t mixtures for several values of g. For the normal mixtures, we estimated parameters using the method described by Basford and McLachlan [29], which is equivalent to applying the EM (expectation-maximization) algorithm [30] to this problem. The ith component pdf of a normal mixture is of the form

$$f_i(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = (2\pi)^{-1}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix, respectively, of the ith component distribution. Thus, letting $\boldsymbol{\phi}$ comprise $\mathbf{p}$, $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_g$, we find that the likelihood function corresponding to a random sample $\mathbf{x}_1,\ldots,\mathbf{x}_n$ from G is proportional to

$$L(\boldsymbol{\phi}) = \prod_{j=1}^{n}\sum_{i=1}^{g} p_i|\boldsymbol{\Sigma}_i|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_i)\}.$$

The likelihood equation,

$$\partial \log L(\boldsymbol{\phi})/\partial \boldsymbol{\phi} = \mathbf{0}, \tag{2}$$

is equivalent to the equations

$$\begin{aligned} \hat{p}_i &= \sum_{j=1}^{n} \hat{w}_{ij}/n, & (3)\\ \hat{\boldsymbol{\mu}}_i &= \sum_{j=1}^{n} \hat{w}_{ij}\mathbf{x}_{ij} / \sum_{j=1}^{n} \hat{w}_{ij}, & (4)\\ \hat{\boldsymbol{\Sigma}}_i &= \sum_{j=1}^{n} \hat{w}_{ij}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)' / \sum_{j=1}^{n} \hat{w}_{ij}, & (5) \end{aligned}$$

for $i = 1, \ldots, g$, where

$$\hat{w}_{ij} = \frac{\hat{p}_i |\hat{\boldsymbol{\Sigma}}_i|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)'\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)\}}{\sum_{t=1}^{g} \hat{p}_t |\hat{\boldsymbol{\Sigma}}_t|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_t)'\hat{\boldsymbol{\Sigma}}_t^{-1}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_t)\}}. \tag{6}$$

The $\hat{w}_{ij}$ are weights such that $\hat{w}_{ij}$ is an estimate of the probability that observation j belongs to component group i. Equations (3)-(6) can be solved iteratively upon first making an initial assignment of observations to groups and supplying an initial estimate of $\boldsymbol{\phi}$ to (6), and then iterating until convergence. The resulting estimate of $\boldsymbol{\phi}$ is a solution to (2) and is thus a local maximum of $L(\boldsymbol{\phi})$. However, it is generally not a global maximum; in fact, (2) has multiple roots, and $L(\boldsymbol{\phi})$ is unbounded so the maximum likelihood estimator of $\boldsymbol{\phi}$ does not exist [31]. Nevertheless, for mixtures of univariate normals it is known that the sequence of roots of (2) corresponding to the largest of the local maxima is consistent, asymptotically normal, and efficient [32], and the same result is widely believed to hold for mixtures of bivariate normals as well. We refer to the root corresponding to the largest of the local maxima as the likelihood-based estimate. To increase the prospects of finding the largest of the local maxima, it is recommended that the iterative solution process begin from several different initial values. The jth observation may be given a final assignment to a group on the basis of the maximum of the converged $\hat{w}_{ij}$ across i.

The normal mixture likelihood-based estimation method just described was carried out for the Carroll County positional error data using the FORTRAN program EMMIX written by D. Peel and G.J. McLachlan, which can be downloaded freely from `http://www.maths.uq.edu.au~gjm/emmix/EMMIX.f`.

To obtain the initial classification of the data needed for starting the estimation algorithm, the data were partitioned randomly into g groups 50 times, and the partition that produced the highest likelihood was adopted as the initial classification. The proportion of observations belonging to the ith group in this initial classification was taken as the initial estimate of p_i, and the sample mean vector and sample covariance matrix of the observations belonging to the ith group were taken an initial estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively.

For the t mixture models, we obtained likelihood-based estimates of parameters using the ECM (expectation-conditional maximization) method described by McLachlan and Krishnan [33]. The ith component pdf of a t mixture is of the form

$$f_i(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i) = \frac{\Gamma(1+\frac{\nu_i}{2})|\boldsymbol{\Sigma}_i|^{-1/2}}{\pi\nu_i\Gamma(\nu_i/2)\{1+(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)/\nu_i\}^{1+\nu_i/2}} \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix, respectively, and ν_i is the degrees of freedom parameter, of the ith component distribution. The degrees of freedom may be viewed as a robustness (to outliers) tuning parameter: a component t pdf with small ν has heavy tails, but as ν tends to infinity the tails become lighter and the corresponding t component pdf tends to a normal pdf. The likelihood function corresponding to a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a g-component t mixture G is then given by

$$L(\boldsymbol{\phi}) = \prod_{j=1}^{n}\sum_{i=1}^{g} p_i f_i(\mathbf{x}_j : \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i),$$

with $f_i(\cdot)$ defined in (7) and with $\boldsymbol{\phi}$ comprising $p_1, \ldots, p_g$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g$, and $\nu_1, \ldots, \nu_g$. Details of the implementation of the ECM estimation algorithm to t mixture models are too lengthy to report here; however, they can be found in [34]. The algorithm was implemented for the Carroll County positional error data using the same program that was used to fit normal mixtures, viz. EMMIX, and the same random grouping scheme used for normal mixtures was used to initially classify the data and obtain initial parameter estimates.

Choosing the number of components

In the previous subsection it was tacitly assumed that the number of components in the mixture distribution was known. While this assumption is appropriate for some applications of mixture models, for example when the subpopulations are males and females or a known number of age classes, it is generally not appropriate for modeling positional errors incurred by geocoding. Thus, the number of components in a mixture distribution for positional errors must be determined using the data at hand. Several methods for accomplishing this have been proposed, ranging from informal graphical techniques to more formal hypothesis testing procedures. Here, we choose the number of components using the *BIC* (Bayesian Information Criterion), a commonly-used model selection method less formal than hypothesis testing but more formal than mere graphical analysis [35]. For a model with k parameters to be estimated, *BIC* is given by

$$BIC = -2 \log L(\hat{\boldsymbol{\phi}}) + k \log n$$

where $L(\hat{\boldsymbol{\phi}})$ is the likelihood function for the n observations, evaluated at the likelihood-based estimator $\hat{\boldsymbol{\phi}}$. *BIC* combines a measure of badness-of-fit, $-2 \log L(\hat{\boldsymbol{\phi}})$, with a measure of model complexity, $k \log n$. When comparing two models, the model with the smaller *BIC* is to be preferred, apart from any other considerations. In the present context, however, we value model parsimony even more highly than usual because of the compelling need for simplicity in measurement-error modeling approaches for handling location uncertainty in spatial analyses. Therefore, although we will use *BIC* as a guide for model selection, we may prefer a model with a slightly larger *BIC* than another if it has many fewer parameters.

Mixture modeling example

We provide the following example to illustrate the mixture model estimation and model selection methodology. Two hundred observations were simulated from a bivariate normal distribution with means $\mu_X = \mu_Y = 0$ (for both variables), variances $\sigma_X^2 = \sigma_Y^2 = 64$, and correlation coefficient $\rho = 0$; and another 200 observations were simulated from a bivariate normal distribution with means $\mu_X = \mu_Y = 10$, variances $\sigma_X^2 = \sigma_Y^2 = 400$, and correlation coefficient $\rho = 0.75$. Each group of observations and their superposition is displayed in Figure 1 (upper panels and lower left panel). Normal mixture models with $g = 1, 2, 3, 4$, or 5 components were fit to these data using EMMIX. Values of *BIC* for these fitted models were 6469, 6387, 6408, 6420, and 6442, respectively. Thus, the two-component model fits best, as was expected. For the two-component model, likelihood-based parameter estimates were as follows:

First component: $\hat{p}_1 = 0.53$, $\hat{\mu}_X = 0.3$, $\hat{\mu}_Y = -0.5$, $\hat{\sigma}_X^2 = 55.7$, $\hat{\sigma}_Y^2 = 60.3$, $\hat{\rho} = 0.01$

Second component: $\hat{p}_2 = 0.47$, $\hat{\mu}_X = 10.9$, $\hat{\mu}_Y = 11.5$, $\hat{\sigma}_X^2 = 446.8$, $\hat{\sigma}_Y^2 = 367.9$, $\hat{\rho} = 0.75$.

These estimates match the true parameter values very well. Finally, the fitted mixture model was used to generate a new set of 400 observations, which are also displayed in Fig. 1 (lower right panel). Upon comparing this display with that for the original set of observations, we see that the fitted model generates data that closely resemble the original simulated data. In this sense, then, the fitted model has excellent predictive power.

Results and Discussion

100%-matched automated geocoding errors

Of the 2354 rural addresses in Carroll County with orthophoto-derived geocodes, 1423 (60.5%) geocoded using the automated method with a 100%-match criterion. The positional errors (which are two-dimensional vectors) associated with these geocodes ranged in length from a minimum of 3 m to a maximum of 2896 m, with a median of 168 m, and are displayed as points in Figure 2. Interestingly, the errors tend to cluster along the N-S and E-W axial directions in such a way that the overall shape of their distribution, apart from a few outliers, resembles a Greek cross (Fig. 2, upper left panel). More errors lie near the center of the cross than near its extremities. Moreover, there is a distinct shift in the mean with respect to the origin along each axial direction: along the E-W axis many more errors occur to the east of zero, while along the N-S axis many more errors occur to the south of zero. Close scrutiny also indicates the existence of two parallel "strands" of errors along each axial direction, which straddle the axes and are likely due to relatively small offsets of residences from street centerlines. Still more interesting features become apparent upon plotting the errors for the 662 addresses on streets running mainly E-W separately from the errors for the 761 addresses on streets running mainly N-S (Fig. 2, upper right and lower left panels). This decomposition shows that while the errors near the cross's center appear to be relatively isotropic, i.e. occurring more or less equally often in all directions, those errors away from the center tend to be aligned with the axial orientation of the street on which the corresponding address lies.

Manual checking of the fifty largest errors revealed that many were attributable to street segments in the TIGER/Line file that had correct street names but incorrect address ranges. Others appeared to be attributable to interpolation errors or possibly house address numbering "errors" (i.e. deviations from the distance-from-intersection rule or some other rule that was used when the houses were originally numbered). These database and procedural errors, in combination with the high degree of rectilinearity of the rural road network in Carroll County, produce the distinctive Greek-cross shape of the empirical distribution of positional errors. Outliers from this overall shape appear to be due to either very large offsets (e.g., one house was nearly 800 m from its corresponding street centerline), incorrect TIGER/Line file geometry, or both.

We do not have a ready explanation for the bias with respect to the origin exhibited by the errors. However, the fact that the mean errors are shifted to the east along E-W streets and south along N-S streets, in tandem with the fact that these directions of shift coincide with the directions in which rural house numbers are ascending, suggest that the explanation has something to do with a systematic interpolation or house numbering procedural error. As a follow-up, we computed the mean error for each individual street and found that these means were consistently, in fact invariably, to the east and south. Thus the bias is pervasive, not limited to merely a few streets.

Owing to the Greek-cross shape of the empirical distribution of the entire set of positional errors, no single bivariate normal or t distribution will fit them well, nor for that matter will *any* elliptical distribution (i.e. a distribution whose contours of equal probability are ellipses). However, the decay in frequency of points with increasing distance from a central location along each axis suggests that a mixture of two or more normal or t distributions, of which at least one is aligned in approximately a N-S direction and at least one other is aligned in approximately an E-W direction, might provide an adequate fit. Consequently, normal and t mixtures with various numbers of components were fit to the errors. Values of *BIC* for each mixture model are given in Table 1a. The results indicate that a three-component mixture fits much better than a two-component mixture, but increasing the number of components beyond three results in marginal improvement in fit. The results also show the t mixture model to be superior to the normal mixture model. In light of these results and taking into account the premium on simplicity in measurement-error models, we would select the three-component t model for these errors.

Likelihood-based estimates of the mean vector and covariance matrix for the three-component t model are given in Table 2a, and Figure 3 depicts 1423 simulated observations from the fitted model. (The number of simulated observations was chosen to match the number of real observations so that plots would be directly comparable.) Upon comparing the lower right panel of Fig. 3 with the upper left panel of Fig. 2, we see that the fitted model reproduces the large-scale features of the positional errors quite well. Furthermore, the parameter estimates and component plots indicate that: (1) the largest component group consists of errors which are mostly "small" (less than 100 m), relatively isotropic, and centered at the origin, but heavy-tailed ($\nu = 1.6$) and thus including some outliers; (2) the other two component groups, comprising relative proportions roughly equivalent to the relative numbers of addresses on N-S and E-W streets, respectively, include many errors of intermediate to relatively large size (> 500 m), which are aligned in the N-S and E-W axial directions, respectively, but are lighter-tailed ($\nu = 6.5$ and $\nu = 19.6$) than the first component and hence relatively devoid of outliers; and (3) the means of the second and third components are several hundred meters to the east and south, respectively, of the origin, which is consistent with the systematic bias in these directions noted previously.

The lower right panel of Fig. 2 displays the "aligned errors," i.e. the errors relative to the axial orientation of the street segment on which the corresponding address lies. Equivalently, the aligned errors are a superposition of the points in the upper right panel and those resulting from a 90-degree counterclockwise rotation of the lower left panel of Fig. 2. Normal and t mixtures were also fitted to the aligned errors. Values of *BIC* and likelihood-based parameter estimates are given in Tables 1b and 2b, respectively. The results suggest that a two-component t mixture fits adequately well; that the first component of this mixture is essentially the same as the first component of the three-component t mixture for the original errors; and that the second component is essentially the combination of the third component and rotated second component of the three-component t mixture for the original errors. In fact, *BIC* for the two-component t mixture for the aligned errors is substantially smaller than *BIC* for the three-component t mixture for the original errors (Table 1), which indicates that accounting for the orientation of the street on which an address lies results in a more parsimonious model with no reduction in model adequacy.

60%-matched automated geocoding errors

Of the 2354 rural addresses in Carroll County with orthophoto-derived geocodes, 1449 (61.6%) geocoded using the automated method with a 60%-match criterion. Note that only 26 more addresses geocoded using this criterion than geocoded using a 100%-match criterion. Figure 4 displays the corresponding positional errors, which range in length from 3 m to 15748 m, with a median of 168 m. The major difference between these errors and those of the 100%-matched data considered previously is the presence of eight extremely large errors, or outliers. The outliers are relatively gregarious, i.e. they tend to occur in clusters. A closer look at the addresses corresponding to the outliers revealed that the extremely large positional errors were due to errors in the TIGER street centerline file such as an incorrect zip code, an address range for a street segment that fails to contain the house number, or a missing street segment. As a consequence of the automated geocoding software's matching algorithm, these errors tended to result in geocodes corresponding to an address with the same house number but lying on a street segment with a different but similar "name," e.g. "120th St" rather than "210th St," or "20th St" rather than "260th St."

Rare, gregarious outliers such as those existing among these data present a severe challenge to any modeling enterprise, including the mixture modeling approach. Consequently, for our purposes here we set these outliers aside and fitted normal and t mixture models to the remaining 1441 60%-matched observations. As with the the 100%-matched data, models were fitted to the original errors and to the errors relative to the alignment of the corresponding street segment. Owing to the very high proportion of observations that the 100%-matched and 60%-matched data have in common, the results of these fits were very similar to those for the 100%-matched data, indeed so much so that we do not display the results.

E911 geocoding errors

The positional errors corresponding to the 2354 E911 geocodes (Fig. 5) ranged in length from 2 m to 974 m, with a median of 44 m. Thus, these errors tend to be considerably smaller than their automated geocoding counterparts. The upper left panel of Fig. 5 shows the errors to be arrayed in a Greek cross-like configuration that appears even more pronounced than was the case for the automated geocoding errors, so likewise a single normal or t distribution will not fit well. But once again there is an attenuation in the frequency of points with increasing distance from a central point along each axis, suggesting that a mixture of two or more normal or t distributions might fit the data well. Moreover, the aforementioned central point of the distribution appears to be at or very close to the origin; there is not a mean shift with respect to the origin along each axis as there was for the automated geocoding errors. Nor do there appear to be "strands" of points straddling, and running parallel to, each coordinate axis, as there were for the automated geocoding errors. However, there are outliers, and there is an interesting effect of orientational alignment: upon plotting the

1116 addresses on streets aligned mainly E-W separately from the 1238 addresses on streets aligned mainly N-S (Fig. 5, upper right and lower left panels), we observe that the errors tend to be aligned *orthogonally* to the orientation of the street on which the corresponding address lies. This is in sharp contrast to the *coincident* alignment of automated geocoding errors with the axial orientation of the street, which we noted previously (Fig. 2).

The orthogonal alignment of E911 errors occurs as a result of offset errors of substantial magnitude, which in turn are due to the definition of the E911 geocode in rural areas as the coordinates of the intersection of the public road and private road leading to the residence, coupled with the approximate perpendicularity (in most cases) of the angle between the public and private road. The outliers, for the most part, correspond to those cases for which the offset is relatively large and the private road meanders in such a way that a hypothetical line segment connecting the residence to the public road-private road intersection is far from being perpendicular.

Normal and t mixture distributions with various numbers of components were fitted to the E911 errors. Values of *BIC* for these fits are listed in Table 1c. On the basis of these values, it appears that a three-component t mixture model provides the best fit; normal models, as well as t models with less than three components, are inadequate. Likelihood-based parameter estimates for the three-component model are given in Table 1c in order of decreasing p_i, and Figure 6 displays 2354 simulated observations from the fitted model. Note that the means of all components lie fairly close to the origin, indicating little systematic bias in the errors. The estimates and component plots reveal that the component comprising the largest proportion (about 52%) consists mostly of relatively small (standard deviation just over 60 m), nearly isotropic errors. The other two components (comprising about 29% and 19% of the errors, respectively) correspond to errors tending to be of larger size (standard deviations of 290 m and 354 m) lying close to the E-W and N-S axial directions, respectively. All three components are quite heavy-tailed, thus outliers occur in all of them. Overall, the simulated data (Fig. 6, lower right panel) again seem to reproduce the observed data (Fig.

5, upper left panel) quite well.

The lower right panel of Fig. 5, which displays the all of the E911 errors relative to the axial orientation of the corresponding street segment, highlights the aforementioned orthogonality of the errors to street orientation. Normal and t mixtures, once again, were fitted to the errors in this plot. Values of *BIC* and likelihood-based parameter estimates are given in Tables 1d and 2d, respectively. According to these results, a two-component t mixture is best-fitting. The component comprising the largest proportion (70%) consists of relatively small errors that are, on average, about twice as large in the orthogonal direction as in the coincident direction. The remaining component consists of much larger errors that average about seven times larger in the orthogonal direction than in the coincident direction. Both components are rather heavy-tailed, indicating that outliers occur regularly for both.

Conclusions

The major question motivating this investigation was whether one could find useful models for the probability distribution of positional errors associated with geocoding, i.e. models that are sufficiently rich to adequately fit various geocoding error datasets yet sufficiently parsimonious to be practical for use as measurement-error models for statistical analysis. The answer to this question, based on our findings, is solidly (though not unequivocally) in the affirmative; and the class of models that seems best suited for the purpose is the class of mixture models of bivariate t distributions. These models can adequately fit such features as clustering along several axial directions, systematic bias in any direction(s), and outliers, all of which occurred in our data; simpler models such as uniform and normal distributions, which have been used previously for positional errors in spatial data, cannot. Moreover, t mixture models are feasible for use with emerging applications of measurement-error methodology to epidemiologic research [19, 22], provided that they consist of very few components. Based on our results and the other published displays of geocoding errors of which we are aware [6, 13], we conjecture that a mixture of three (two) t distributions will usually be sufficient for errors (aligned errors) associated with 100%-matched automated geocoding and E911 geocoding, but additional investigations in other places are needed to substantiate this. Regions with less rectilinear road networks than the Carroll County network may sometimes require fewer components, as the errors on such networks are less likely to exhibit clustering in the E-W and N-S axial directions; a case in point is displayed in [6]. In some cases a single t distribution may even suffice.

The one situation we encountered in which mixture models of t distributions proved to be less than fully successful occurred with automated geocoding errors for which an address-matching threshold of less than 100% was used. In this situation, a few small clusters of extremely large errors occurred. Such errors are difficult to model parsimoniously and, regardless of how they are modeled, will weaken the conclusions made from subsequent statistical inferences using measurement-error methodology. Consequently, we recommend using only 100%-matched addresses for spatial epidemiologic analyses.

Our investigation indicated that t mixture models were equally useful for 100%-matched automated geocoding errors and E911 geocoding errors, despite some differences in their distinctive features. In particular, t mixtures were able to accommodate the difference in the major axis of error alignment relative to the alignment of the corresponding street (parallel for automated geocoding, perpendicular for E911 geocoding). The error distributions associated with other geocoding methods may have their own distinctive features (see [6], for example, for a graphical display of errors incurred by parcel address-matching), and it remains to be seen whether t mixtures are as successful for them.

Further investigation is currently underway to determine if t mixture models are as useful for positional errors corresponding to non-rural addresses as they appear to be for rural address positional errors and, if so, how the components might differ from those for rural addresses. Results from previous studies of positional errors for datasets combining both rural and non-rural addresses [41, 6, 11, 12] suggest strongly that component variances will be smaller for non-rural addresses, but we refrain from predicting how many components may be needed and whether they will prove to be heavy-tailed, mean-shifted away from the origin, etc.

In focusing our attention on geocoding errors, we have ignored the fact that for many studies, automated geocoding is incomplete; that is, not all addresses can be assigned point-level spatial coordinates by the software. In fact, it is common in practice for 20% or even as many as 40% of subjects' addresses to fail to geocode using standard software and street files. For example, Gregorio et al. [36] and Oliver et al. [37] present public health studies in which 14% and 26%, respectively, of the addresses in their datasets could not be assigned a point location via automated geocoding, and for our exclusively rural address dataset this figure was even higher (38%). A statistical analysis based on only the observations that geocode is subject to selection bias [38, 37]. However, there is virtually always a reliable coarse (areal-level) measurement, e.g. a zip code, associated with each observation that fails to geocode. Coarse locational data may be combined with the observed point-level data to make valid statistical inferences in the presence of geographic bias via either (a) a coarsened-data maximum likelihood estimation procedure [39], or (b) imputation of a surrogate point location (such as that of a randomly selected event within the same zip code) for the addresses that do not geocode [40]. Fully satisfactory inference procedures for data whose point locations are ascertained by automated geocoding may require that an inference procedure developed for use with incompletely geocoded data be combined with modifications to account for positional errors.

Authors' contributions

DLZ conceived of this study and drafted the majority of the manuscript. DLZ also directed, and XF performed, the statistical analysis. SM performed the automated geocoding and oversaw the orthophoto geocoding of the Carroll County data and contributed to the writing of the Methods section. GR contributed to the writing of several sections.

Acknowledgements

The work of the authors was supported by Centers for Disease Control and Prevention (CDC)

Grant Number 3 R01 EH000056-01S1 with the Iowa Department of Public Health (IDPH) and Contract Number 5886CAR02 between the IDPH and the University of Iowa. The views expressed are solely those of the authors and do not represent the views of CDC or IDPH. We thank Carl Wilburn, GIS Coordinator for Carroll County, Iowa for providing address data and E911 geocodes for Carroll County.

References

1. Thomas RW (ed.): *Spatial Epidemiology*, London Papers in Regional Science 21. London: Pion Ltd.; 1990.

2. Elliott P, Wakefield JC, Best NG, Briggs DJ: *Spatial Epidemiology: Methods and Applications*. Oxford, UK: Oxford University Press; 2000.

3. Lawson AB: *Statistical Methods in Spatial Epidemiology*. New York: John Wiley & Sons; 2001.

4. Waller LA, Gotway CA: *Applied Spatial Statistics for Public Health Data*. Hoboken, New Jersey: John Wiley & Sons; 2004.

5. Jacquez GM: **Current practices in the spatial analysis of cancer: flies in the ointment.** *International Journal of Health Geographics* 2004, **3**:22.

6. Cayo MR, Talbot TO: **Positional error in automated geocoding of residential addresses.** *International Journal of Health Geographics* 2003, **2**:10.

7. Yang DH, Bilaver LM, Hayes O, Goerge R: **Improving geocoding practices: Evaluation of geocoding tools.** *Journal of Medical Systems* 2004, **28**:361-370.

8. Kravets N, Hadden WC: **The accuracy of address coding and the effects of coding errors.** *Health & Place*, in press.

9. Dearwent SM, Jacobs RR, Halbert JB: **Locational uncertainty in georeferencing public health datasets.** *Journal of Exposure Analysis & Environmental Epidemiology* 2001, **11**:329-334.

10. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW: **On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research.** *American Journal of Public Health* 2001, **91**:1114-1116.

11. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL: **Positional accuracy of geocoded addresses in epidemiologic research.** *Epidemiology* 2003, **14**:408-412.

12. Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G: **Accuracy and repeatability of commercial geocoding.** *American Journal of Epidemiology* 2004, **160**:1023-1029.

13. Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss, G: **Accuracy of commercial geocoding: assessment and implications.** *Epidemiologic Perspectives and Innovations* 2006, **3**:8.

14. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P: **Positional accuracy of two methods of geocoding.** *Epidemiology* 2005, **16**:542-547.

15. Waller LA: **Statistical power and design of focused clustering studies.** *Statistics in Medicine* 1996, **15**:765-782.

16. Jacquez GM, Waller LA: **The effect of uncertain locations on disease cluster statistics.** In *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing.* Edited by Mowrer HT, Congalton RG. Chelsea, Michigan: Arbor Press; 2000:53-64.

17. Zimmerman DL: **Statistical methods for incompletely and incorrectly geocoded cancer data.** In *Cancer Geocoding: Ensuring Fitness for Use*. Edited by Rushton G, Armstrong MA, Gittler J, Greene B, West M, Zimmerman DL. CRC Press; in press.

18. Burra T, Jerrett M, Burnett RT, Anderson M: **Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario.** *The Canadian Geographer* 2002, **46**:160-171.

19. Diggle PJ: **Point process modelling in environmental epidemiology.** In *Statistics for the Environment*. Edited by Barnett V, Turkman KF. New York: John Wiley & Sons; 1993:89-110.

20. Gabrosek J, Cressie N: **The effect on attribute prediction of location uncertainty in spatial data.** *Geographical Analysis* 2002, **34**:262-285.

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

22. Zimmerman DL, Sun P: **Estimating spatial intensity and variation in risk from locations subject to geocoding errors.** *Technical report #363, Department of Statistics and Actuarial Science, University of Iowa*; 2006:1-19. [http://www.stat.uiowa.edu/techrep/tr363.pdf]

23. Barber JJ, Gelfand AE, Silander JA: **Modeling map positional error to infer true feature location.** *Canadian Journal of Statistics*, in press.

24. Everitt BS, Hand DJ: *Finite Mixture Distributions*. London: Chapman and Hall; 1981.

25. Titterington DM: *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons; 1985.

26. McLachlan GJ, Basford KE: *Mixture Models*. New York: Marcel Dekker; 1988.

27. ArcGIS 9: *Geocoding Rule Base Developers Guide*. Redlands, California: Earth Sciences Research Institute; 2003.

28. Postal Addressing Standards-Publication 28: United States Postal Service; 2000. Accessed from `http://pe.usps.com/cpim/ftp/pubs/Pub28/pub28.pdf` on Sept. 23, 2006.

29. Basford KE, McLachlan GJ: **Likelihood estimation with normal mixture models.** *Applied Statistics* 1985, **34**:282-289.

30. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm (with discussion).** *Journal of the Royal Statistical Society Series B* 1977, **39**:1-38.

31. Kiefer J, Wolfowitz J: **Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters.** *Annals of Mathematical Statistics* 1956, **27**:887-906.

32. Kiefer NM: **Discrete parameter variation: efficient estimation of a switching regression model.** *Econometrica* 1978, **46**:427-434.

33. McLachlan GJ, Krishnan T: *The EM Algorithm and Extensions*. New York: John Wiley & Sons; 1997.

34. Peel D, McLachlan GJ: **Robust mixture modelling using the t distribution.** *Statistics and Computing* 2000, **10**:339-348.

35. Burnham KP, Anderson DR: *Model Selection and Multi-Model Inference*. New York: Springer-Verlag; 1998.

36. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ: **Subject loss in spatial analysis of breast cancer.** *Health & Place* 1999, **5**:173-177.

37. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW: **Geographic bias related to geocoding in epidemiologic studies.** *International Journal of Health Geographics* 2005, **4**:29.

38. Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, Savitz DA, Herring AH: **Comparison of residential geocoding methods in population-based study of air quality and birth defects.** *Environmental Research* 2006, **101**:256-262.

39. Zimmerman DL: **Estimating spatial intensity and variation in risk from locations coarsened by incomplete geocoding.** *Technical report #362, Department of Statistics and Actuarial Science, University of Iowa*; 2006:1-28. [http://www.stat.uiowa.edu/techrep/tr362.pdf]

40. Boscoe F: **The science and art of geocoding: Tips for improving match rates and handling unmatched cases in analysis.** In *Cancer Geocoding: Ensuring Fitness for Use.* Edited by Rushton G, Armstrong MA, Gittler J, Greene B, West M, Zimmerman DL. CRC Press; in press.

41. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA: **Geocoding addresses from a large population-based study: lessons learned.** *Epidemiology* 2003, **14**:399-407.

Table 1: Bayesian Information Criteria (*BIC*) for normal and t mixture models. Models with several different numbers of components, were fitted to the following four error datasets: (a) 100%-matched automated geocoding positional errors; (b) 100%-matched automated geocoding positional errors aligned with axial direction of corresponding street segment; (c) E911 positional errors; (d) E911 positional errors aligned with axial direction of corresponding street segment.

Error dataset	Distribution	Number of Components	*BIC*
(a)	Normal	1	48103
	Normal	2	45851
	Normal	3	45236
	Normal	4	45124
	t	1	46083
	t	2	45358
	t	3	45056
	t	4	45042
(b)	Normal	1	46422
	Normal	2	44809
	Normal	3	44597
	Normal	4	44557
	t	1	45659
	t	2	44538
	t	3	44516
	t	4	44459
(c)	Normal	1	67174
	Normal	2	63174
	Normal	3	62710
	Normal	4	62446
	t	1	62841
	t	2	62345
	t	3	62219
	t	4	62230
(d)	Normal	1	64227
	Normal	2	61360
	Normal	3	61101
	Normal	4	61059
	t	1	61092
	t	2	60980
	t	3	60982
	t	4	60994

Table 2: Likelihood-based parameter estimates for the best-fitting models. Models and the datasets to which they were fitted are: (a) the three-component t mixture model for the 100%-matched automated geocoding positional errors; (b) the two-component t mixture model for the 100%-matched automated geocoding positional errors aligned with axial direction of corresponding street segment; (c) the three-component t mixture model for the E911 positional errors; (d) the two-component t mixture model for the E911 positional errors aligned with axial direction of corresponding street segment. Means are denoted by μ_X and μ_Y, standard deviations by σ_X and σ_Y, correlation coefficient by ρ, and degrees of freedom by ν. Units of measurement for means and standard deviations are meters.

Error dataset	Component	Proportion	μ_X	μ_Y	σ_X	σ_Y	ρ	ν
(a)	1	0.571	-12.1	-10.7	61.6	54.1	-0.05	1.6
	2	0.253	-4.7	-350.0	75.9	550.0	0.18	6.5
	3	0.176	352.8	-12.6	540.3	84.9	-0.03	16.7
(b)	1	0.560	-0.8	-14.2	39.4	75.9	0.06	1.8
	2	0.440	372.1	-6.7	523.6	90.3	-0.10	5.9
(c)	1	0.519	4.9	-5.4	62.3	60.8	-0.10	1.8
	2	0.292	-13.6	-35.0	289.1	54.9	-0.14	2.4
	3	0.189	14.9	-10.2	62.1	354.4	0.14	2.4
(d)	1	0.700	5.9	-4.3	47.0	100.7	0.06	1.8
	2	0.300	29.3	-6.2	62.1	419.5	0.16	3.0

Figure 1: Scatterplot of simulated data from two-component bivariate normal mixture model. Upper left panel: 200 observations from first component; Upper right panel: 200 observations from second component; Lower right panel: Superposition of upper right panel lower left panel; Lower right panel: a new simulation of 400 observations from the two-component normal mixture model fitted to the data from the original simulation.

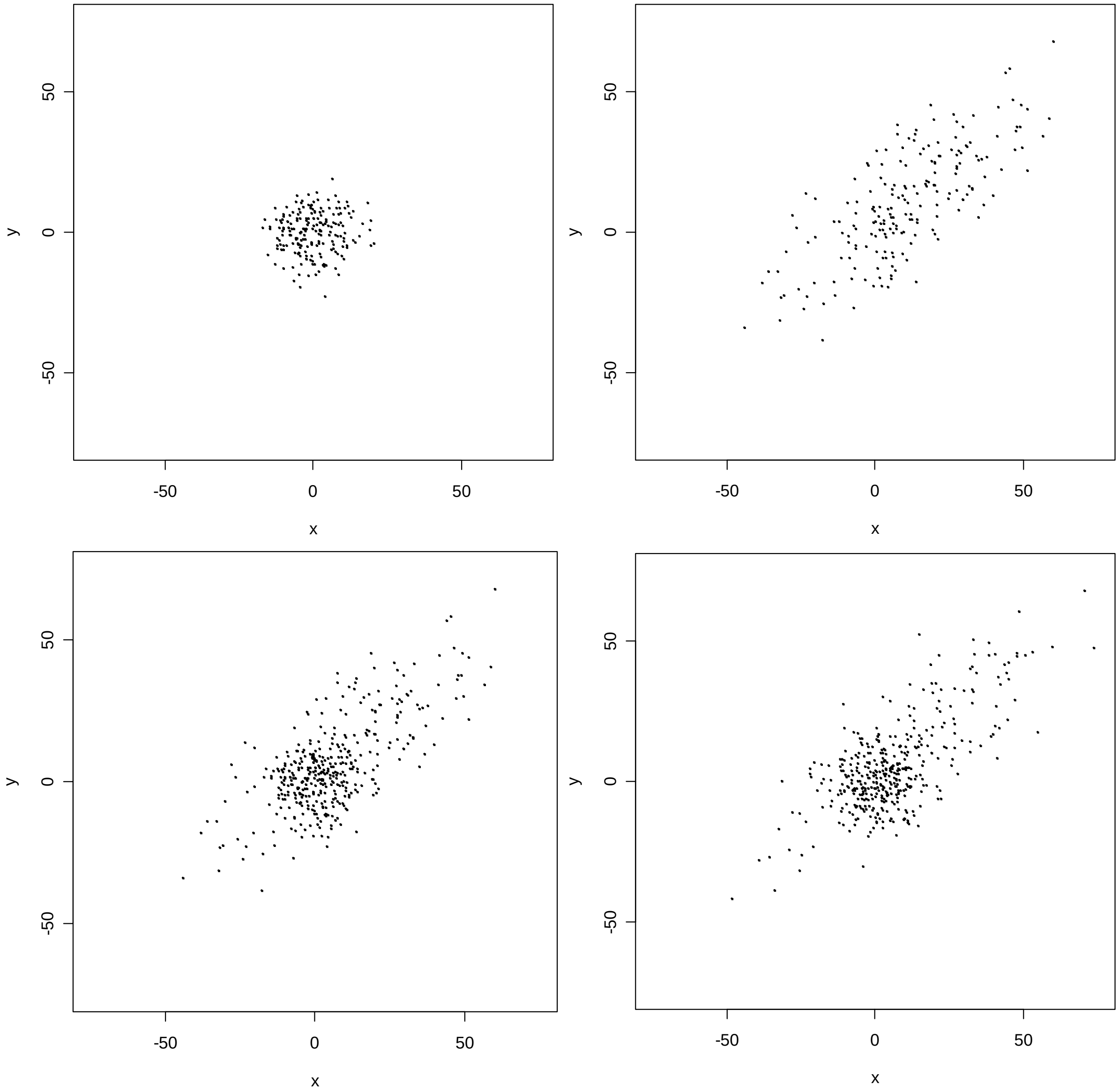


Figure 2: Scatterplot of positional errors (in meters) for the 100%-matched automated geocodes. Upper left panel: Complete data; Upper right panel: errors for addresses on streets aligned E-W; Lower left panel: errors for addresses on streets aligned N-S; Lower right panel: Superposition of upper right panel and 90-degree counterclockwise rotation of lower left panel.

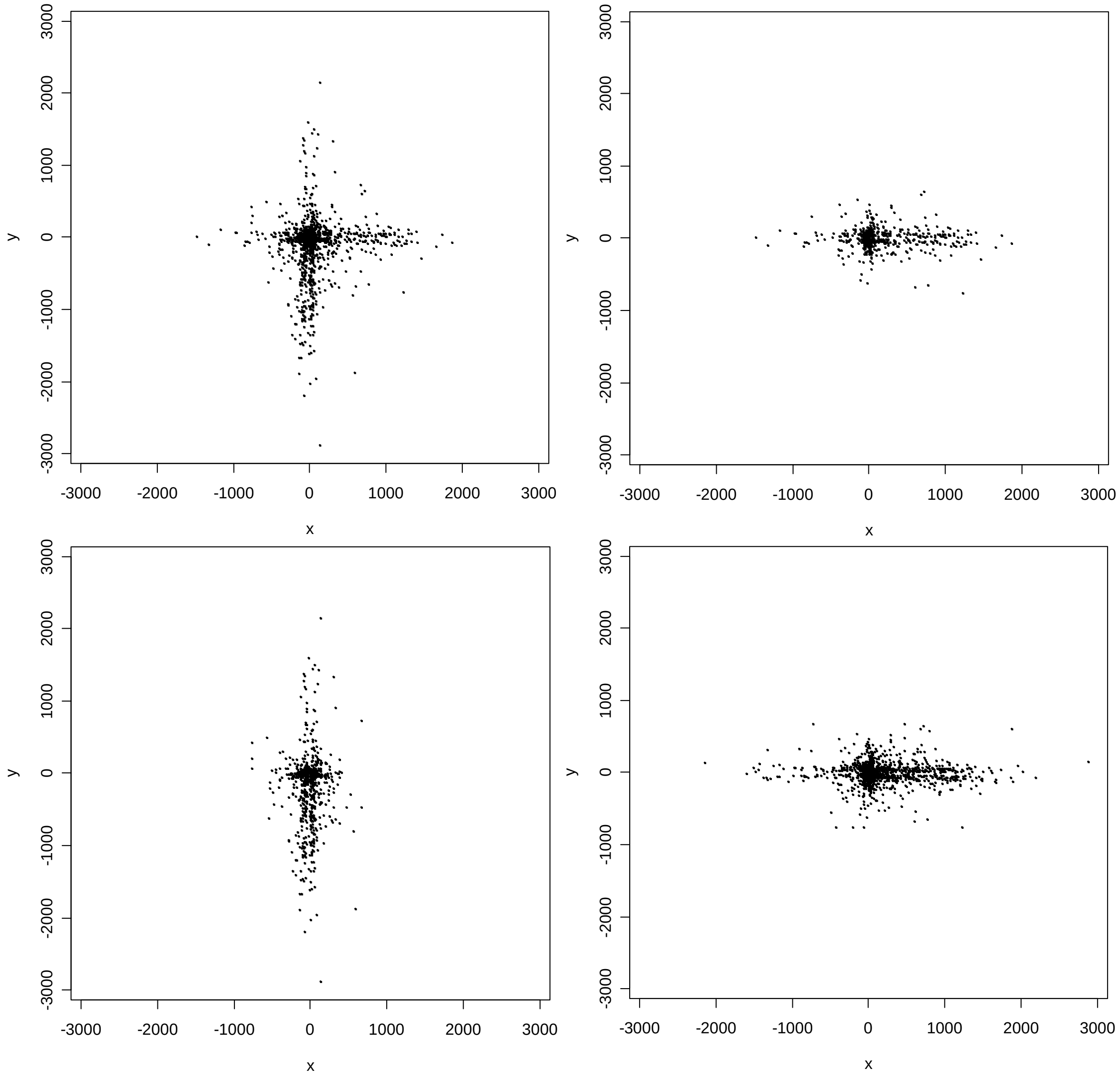


Figure 3: Simulated data from the fitted three-component t mixture distribution for the 100%-matched automated geocoding errors. The upper left panel, upper right panel, and lower left panel correspond to components in order of decreasing p_i; and the lower right panel is their superposition.

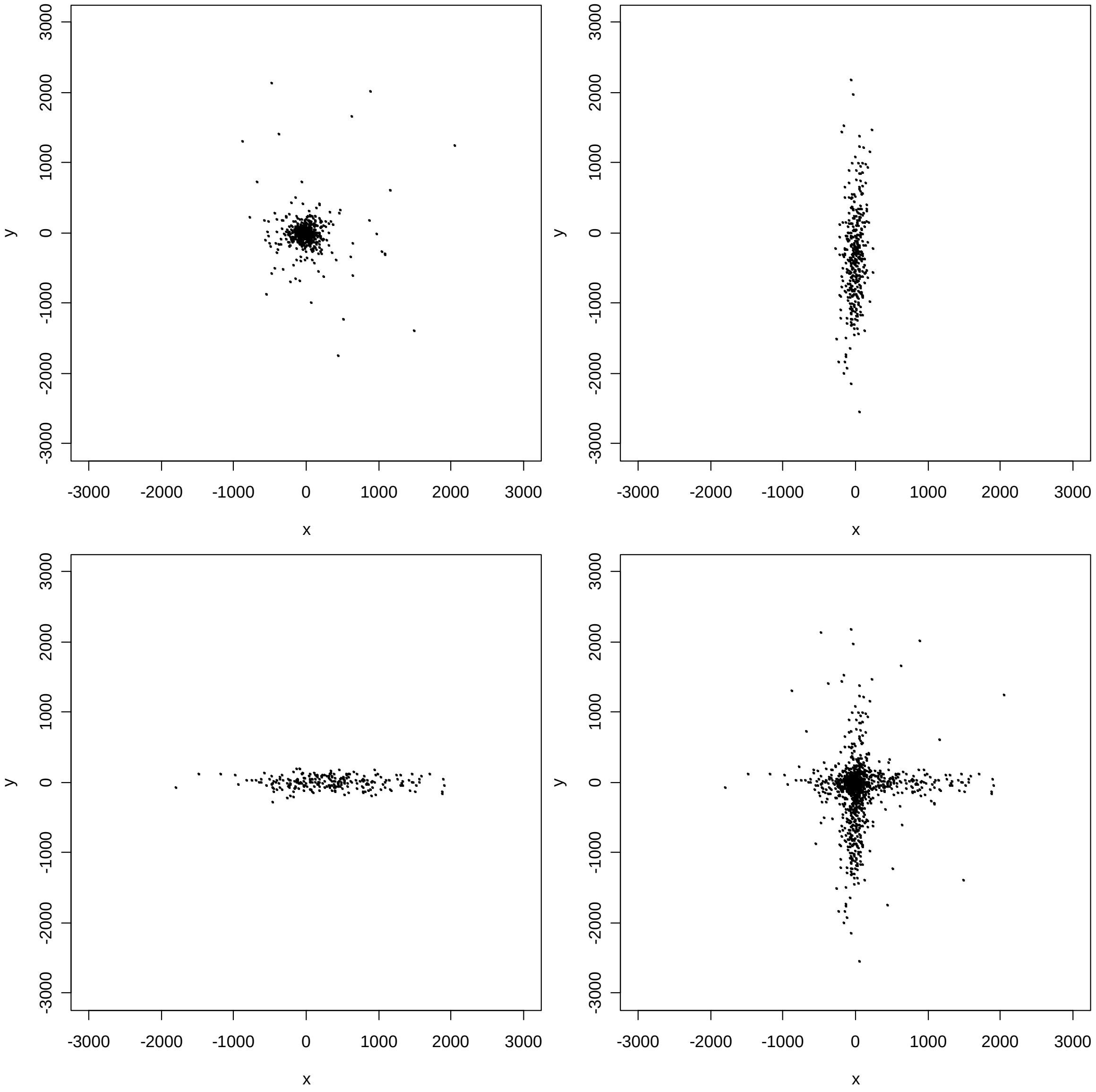


Figure 4: Scatterplot of the positional errors (in meters) for the 60%-matched automated geocodes. Left panel: Complete data; Right panel: Data with eight outliers removed.

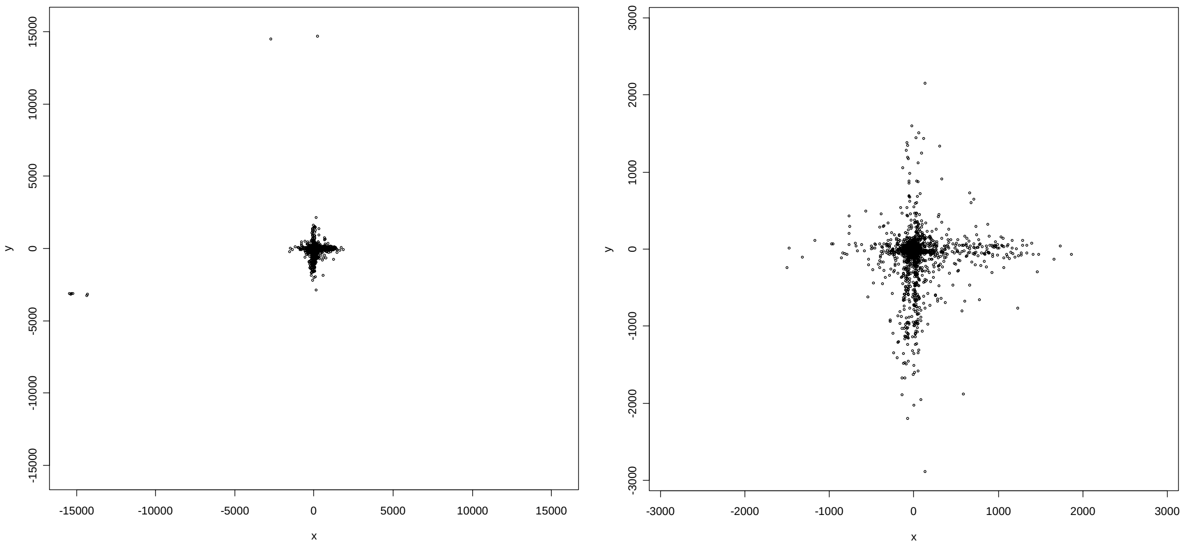


Figure 5: Scatterplot of the positional errors (in meters) for the E911 geocodes. Upper left panel: Complete data; Upper right panel: errors for addresses on streets aligned E-W; Lower left panel: errors for addresses on streets aligned N-S; Lower right panel: Superposition of upper right panel and 90-degree counterclockwise rotation of lower left panel.

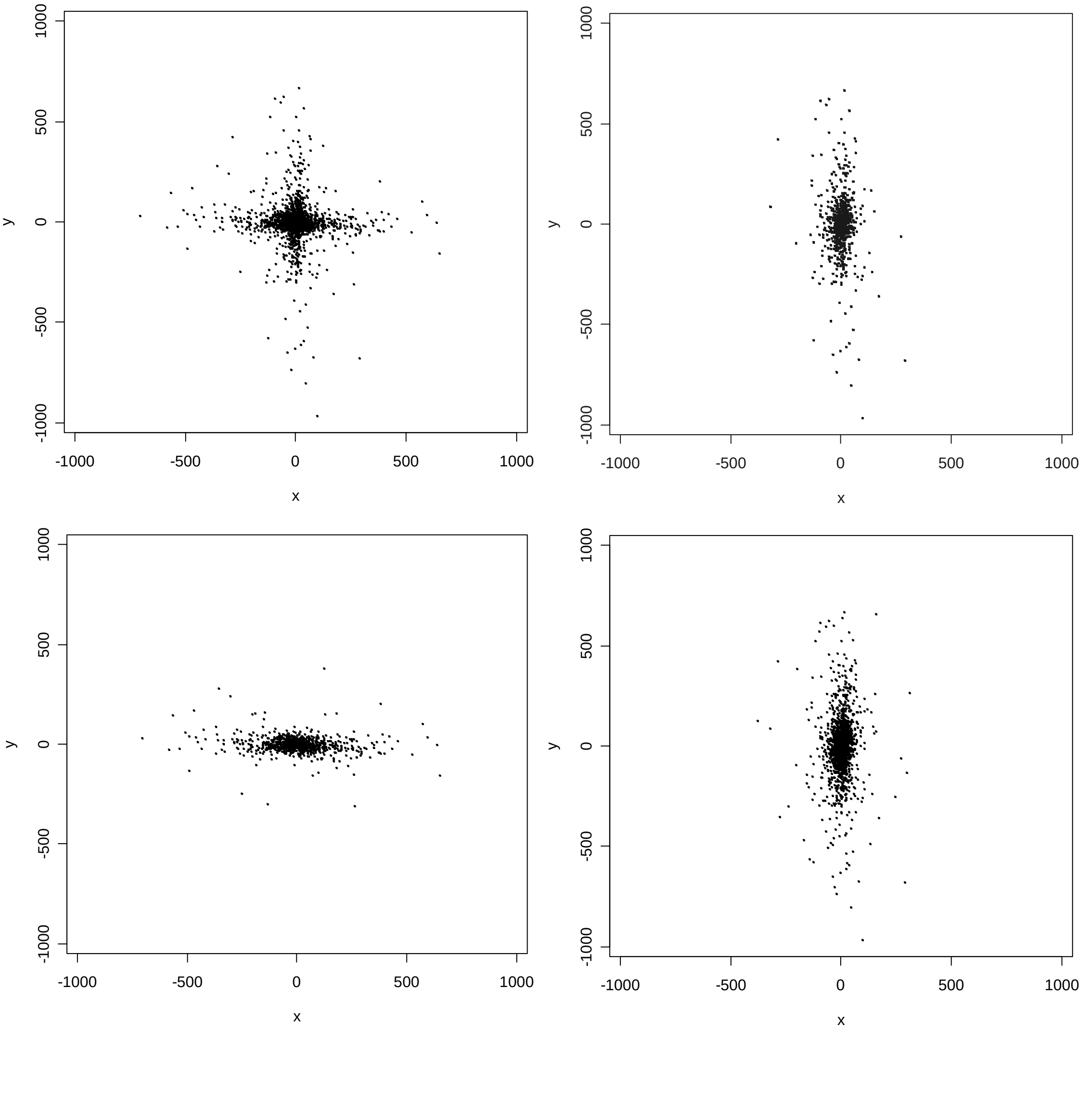
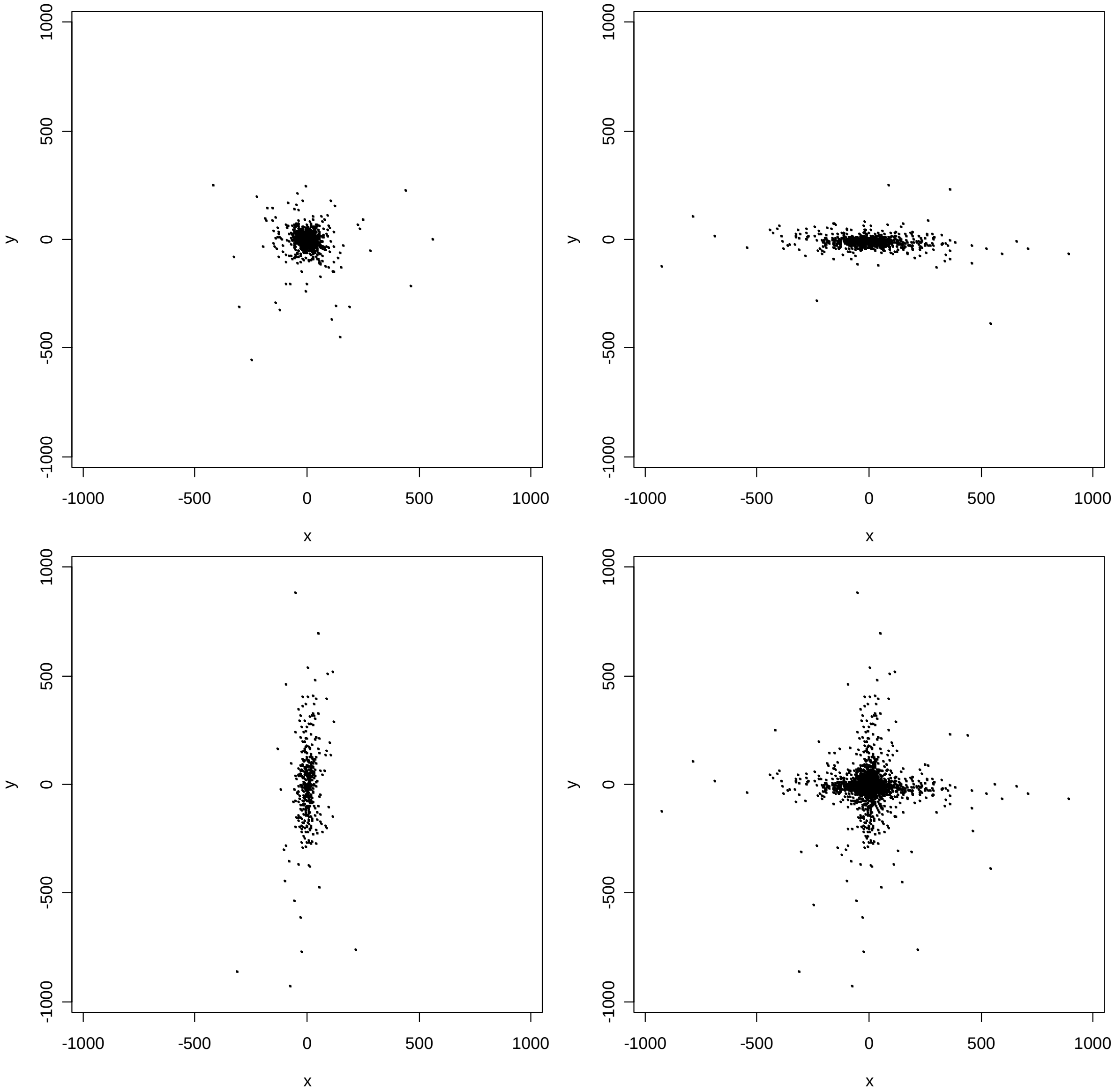


Figure 6: Simulated data from the fitted three-component t mixture distribution for the E911 errors. The upper left panel, upper right panel, and lower left panel correspond to components in order of decreasing p_i; and the lower right panel is their superposition.



y
x
1000
500
0
-500
-1000
-1000
-500
0
500
1000