Popul Environ (2015) 37:83–98
DOI 10.1007/s11111-014-0225-0

ORIGINAL PAPER

"Scrambling" geo-referenced data to protect privacy induces bias in distance estimation

Noam Elkies · Günther Fink · Till Bärnighausen

Published online: 3 February 2015

Abstract Privacy concerns regarding the sharing of spatially referenced household data have induced researchers and survey agencies to "scramble" geographic information by adding random spatial errors to true location coordinates. In this paper, we prove mathematically that the addition of random noise leads to a systematic overestimation of distances between households and access points of interest. We illustrate this average distance bias as well as the attenuation bias generated by random spatial errors using data on household and health facility location from a Health and Demographic Surveillance Site in rural South Africa. Given the large overall biases observed, we argue that the use of scrambled spatial data for policy making or empirical work is generally not advisable, and that alternative methods of protecting data confidentiality should be used to ensure the usability of spatial data for quantitative analysis.

Keywords Coordinate scrambling · Coordinate displacement · Data privacy · Demographic and Health Surveys (DHS) · Confidentiality · GPS

N. Elkies
Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

G. Fink (✉) · T. Bärnighausen
Department of Global Health and Population, Harvard School of Public Health, 665 Huntington Avenue, Building I, Room 1211, Boston, MA 02115, USA
e-mail: gfink@hsph.harvard.edu

T. Bärnighausen
Wellcome Trust Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Mtubatuba, South Africa

Introduction

The rapid growth in the availability of household data containing Geographical Information System (GIS) coordinates has opened new venues to researchers interested in the complex interactions between space, human behavior, and outcomes (Arcury et al. 2005; Cooke et al. 2010; Seiber and Bertrand 2002; Tanser et al. 2009). The availability of geo-referenced household data allows researchers to study the density and spatial distribution of specific features of populations; it also allows researchers to assess the extent to which specific behaviors, such as school attendance or health care seeking, depend on spatial location and distance to public infrastructure like schools or hospitals (Kaplan and Hegarty 2006; Kyei et al. 2012; Lohela et al. 2012).

One of the most critical challenges faced by researchers working with geo-referenced household data is confidentiality and privacy protection (Kamel Boulos et al. 2009; O'Brien and Yasnoff 1999; Onsrud et al. 1994). Given that Global Positioning System (GPS) coordinates generally lie within a 10-m radius of the true location (Schwieger 2003), households and household members could be identified if their GPS coordinates were publicly available. Even in densely populated urban areas, relatively few households are within a 10-m radius of a given coordinate and even fewer may match all the other household characteristics included in the data set, including income, family size, or occupation. Identification of households is of particular concern if GPS coordinates are linked to sensitive data on household members, such as household members' HIV status and sexual behavior. Allowing the identification of survey respondents constitutes a clear violation of the confidentiality that researchers typically guarantee as part of the consent process preceding a subject's enrollment in a research study. If individuals with stigmatized traits or particular vulnerabilities can be identified, a range of harms and undesirable consequences may result, such as harassment, social exclusion, or crime (Hyman 2000).

Several approaches have been developed to address these data confidentiality concerns. Complete non-disclosure, restricted disclosure, and coordinate scrambling are the most common approaches (Golden et al. 2005). Other possible approaches include aggregating high-resolution spatial data to larger administrative units, as well as the use of software agents that allow researchers to analyze fully identified data without being able to physically access the underlying identifiable information (Kamel Boulos et al. 2006).

Complete non-disclosure essentially implies destroying the original household location information upon completion of the study. The approach is easy and highly effective for ensuring data confidentiality, but also implies the systematic destruction of potentially valuable information. *Partial disclosure* is the approach currently taken by some of the larger European and US household surveys, such as the Health, Ageing and Retirement (SHARE) surveys (Borsch-Supan et al. 2013; Linardakis et al. 2013) or the National Longitudinal Study of Youth (Center for Human Resource Research 1997). In these surveys, GPS or address data are collected, but are not part of the data sets that are made publicly available. Upon request, GPS data are selectively made available to researchers with appropriate

scientific reasons for using this data. To minimize the risk of data leakage, researchers generally need to sign strict data confidentiality agreements and may also be required to come to specific data centers to physically access the data. Partial disclosure has two main disadvantages: First, setting up confidentiality agreements with a potentially large number of researchers and research institutions often requires substantial human and legal resources; second, traveling to a data center may be impossible for many researchers because of financial and time constraints.

To avoid these challenges, large data collection operations, such as the Demographic and Health Surveys (DHS) and many Health and Demographic Surveillance Systems (HDSS), use a third approach, which is loosely referred to as "coordinate scrambling" (ICF International 2012). Similar to the "selective availability" program run by the US military until 2000 (National Archives and Records Administration 1996), a random noise vector is added to each coordinate to "mask" its true location, and the resulting "scrambled" coordinate is then made available to the public. In the case of the DHS, the scrambled coordinate falls within a 2-km radius of the original coordinate in urban areas and within a 5-km radius in rural areas.[1] Both scrambling radii were chosen with the objective to make identifying households sufficiently difficult, which essentially means making sure that sufficiently large number of households fall within the chosen radius. This could in theory be achieved by choosing the scrambling radius as a function of local population density. In practice, local population density data are not always available; as a result, large survey operations like the DHS simply use a smaller radius for the typically more densely populated urban areas and a larger radius for rural areas.

Conceptually, the idea of scrambling seems attractive, since scrambling implies protecting data confidentiality, while allowing researchers to work with collected geographical data. In practice, however, neither the theoretical nor the empirical implications of scrambling are well understood. To address this knowledge gap, we formally introduce the concept of scrambling in this paper and mathematically assess implications of scrambling for estimates of average distance as well as estimates of the relationships between distance and other variables of interest. We prove mathematically that scrambling systematically biases the distance estimates between one point whose true coordinates are known to the analyst (e.g., a healthcare clinic) and another point whose true location has been disguised through scrambling (e.g., a household). To illustrate the resulting biases empirically, we scramble true GPS data from a demographic surveillance site (DSS) in rural South Africa and then show the resulting distance and regressions biases in a range of study settings.

The effect of scrambling on average observed distances and distance variation

Let us define a point X as a reference or access point of interest and d as the distance of interest between X and some given household location h. One may think of X as

[1] See http://www.measuredhs.com/What-We-Do/GPS-Data-Collection.cfm for details.

the nearest fast food location, shop selling cigarettes, or healthcare clinic. For simplicity, but without loss of generality, we assume that the access point X is located at the origin, so that the vector from X to each household h is just h itself and has distance $|h|$.

To derive the bias in the average distances computed based on scrambled data, we assume that the true location of h is perturbed by adding a random noise vector v drawn from some centrally symmetric distribution as illustrated in Fig. 1. "Centrally symmetric" requires that for any region R the probability that $v \in R$ is the same as the probability that $-v \in R$. If the random noise vector v is drawn from such a centrally symmetric distribution, the expected value of the noise vector is zero. This means that there is no systematic bias in either direction, so that the perturbed vector $h + v$ has the same expected position as the corresponding vector h.

Under this rather general assumption, we demonstrate that the following proposition is true:

Proposition 1 *For any vector h and a randomly added scrambling vector v the following must always be true:*

(i) *The expected value of the scrambled distance from the point of interest, $\langle |h + v| \rangle$, always exceeds the true distance $|h|$ as long as v is not limited to the line segment between $-h$ and $+h$.*

(ii) *Adding a noise vector v always increases the expected square distance $\langle |h + v|^2 \rangle$, and the difference $\langle |h + v|^2 \rangle - |h|^2$ equals the mean square $\langle |v|^2 \rangle$ of v.*

(iii) *The expected bias between the true and the observed distance is bounded by $\langle |v|^2 \rangle / (2|h|)$; thus, as long as $|v| \leq \gamma$ for some radius γ, then, as long as the distribution of v is symmetric under v, we have $\left\langle |h + v|^2 \right\rangle - |h|^2 \leq \frac{\gamma^2}{2|h|}$.*

Part (i) of Proposition 1 states that the average (expected) scrambled distance is *strictly larger* than the true distance for any two-dimensional error. The magnitude of this bias increases with the maximum scrambling radius and decreases with the true distance as shown in Part (iii) of the proposition. Intuitively, adding two-dimensional noise terms biases the average distance due to the nonlinear relation defined in Pythagoras' Theorem; the (linear) average of the scrambled distances turns out to be systematically larger than the actual true distance. Part (ii) of the proposition is more straightforward; given that the two vectors of interest are by assumption independent, the total variation in scrambled distance can be directly decomposed in the true variation in distance and the average variation generated by the scrambling error.

The full mathematical proof of Proposition 1 is available in "Appendix".

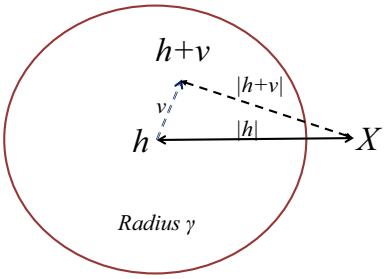

Fig. 1 Model setup

Empirical implications

Proposition 1 has two main implications for empirical analysis. First, and most importantly, any population-based estimate of average distance based on scrambled data will display a systematic upward bias. This overestimation of true distances may have undesirable consequences if estimated distances are used for policy. Assume, for instance, that the government wants to know the average distance children travel to school or the fraction of individuals living outside a given distance to a health facility. Further, assume that the coordinate of the school or health facility is precisely known, but that all coordinates of households have been scrambled. If the government uses these scrambled coordinates to calculate distances, the average observed distance will be strictly larger than the average true distance, so that the fraction of individuals living beyond a given distance of interest will generally be overestimated.

The second major issue when working with scrambled data directly links to the statistical literature on measurement error in variables starting with the seminal work by Spearman (1904). As shown in part (ii) of Proposition 1, the addition of random spatial noise essentially implies adding measurement error to the variable of interest. Since the measurement error is orthogonal to the true distance by construction, the classical-errors-in-variables (CEV) case will arise. In a standard ordinary least squares regression (OLS) framework, the probability limit of the coefficient estimated for the distance variable of interest is given by

$$p\ \lim\left(\hat{\beta}_{\mathrm{OLS}}\right) = \beta \frac{\sigma_h^2}{\sigma_h^2 + \sigma_v^2},$$

where β is the true coefficient of interest and σ_h^2, σ_v^2 correspond to the variance in the true distance and the scrambling error, respectively (Wooldridge 2002, 2003). The greater the variance in the random noise term, the closer the estimated slope moves toward zero; this effect is generally referred to as "regression dilution," "attenuation," or "attenuation bias" following the original work by Spearman.

To illustrate the importance of scrambling biases in practical application, we use data from the Wellcome Trust Africa Centre for Health and Population Studies (Africa Centre) in rural South Africa. As described in further detail in Tanser et al. (2008), the Africa Centre surveillance was launched in 2000 and longitudinally tracks

demographic and health outcomes for all individuals who reside in a geographically contiguous demographic surveillance area covering a total of 438 km^2. The area is mostly rural and densely populated with about 25 households per square kilometer. As of June 2013, the site covers about 90,000 individuals. Figure 2 shows the spatial distribution of households and primary healthcare clinics in the area.

To illustrate the effects of scrambling on estimation, we assume a simple population model, where the outcome of interest y for an individual i (such as school attendance, antenatal clinic attendance, HIV antiretroviral treatment uptake) is a linear function of distance and random error term, such that

$$y_i = \alpha + \beta \text{Dist}_i + \varepsilon,$$

where Dist is distance in kilometers and ε is a randomly distributed error term.

We start our simulations with the basic scenario outlined in the theoretical model and illustrated in Fig. 1, with one specific reference point and a given scrambling radius. In the DSS data, the true coordinates of both the households and the reference points are known. We can thus compare actual distances to the ones observed in a setting where the household coordinates have been scrambled. To evaluate the impact of scrambling on regression analysis, we assume a simple data generating process, where some generic outcome variable y is a linear function of the true distance with a stochastic error terms as described above.

We simulate three different scenarios: a scenario with a very close reference point (inside the demographic surveillance area), a scenario with a mid-range reference point (20 km), and a scenario with a more distant reference point (50 km). For each scenario, we take the true coordinates of all 16,309 households shown in Fig. 2, generate a dependent variable as a linear function of the true distance, and then run 1,000 simulations with scrambled data. Each iteration of the simulation proceeds in three steps: In the first step, we add a random (scrambling) error between 0 and the chosen radius to each of the original household coordinates; in the second step, we compute the Euclidean distance between the scrambled household coordinates and the reference point of interest, and in the last step, we run a regression using the scrambled rather than actual distance as explanatory variables. We store the average distances and regression coefficients obtained in each iteration of the simulation and then compare them to the true values of both variables.

Figure 3 summarizes the results from these simulation models and illustrates the general relation between scrambling noise, observed distances, and expected regression point estimates. More scrambling noise unambiguously increases the expected average distance (bias) as well as the attenuation bias in regressions, while more distant reference points reduce the biases observed. Assuming an average distance from the household to the nearest school or clinic of <10 km (scenario 1), and a scrambling radius of 5 km recommended for rural areas, the average distance is overestimated by 0.51 km, which corresponds to an upward bias of about 5 %. The bias is much larger in regression models, where the average estimated distance coefficients is 36 % smaller than then true effect.

The situation is further complicated when subjects have a more complex choice set (such as multiple schools or clinics) to choose from. Often researchers may wish

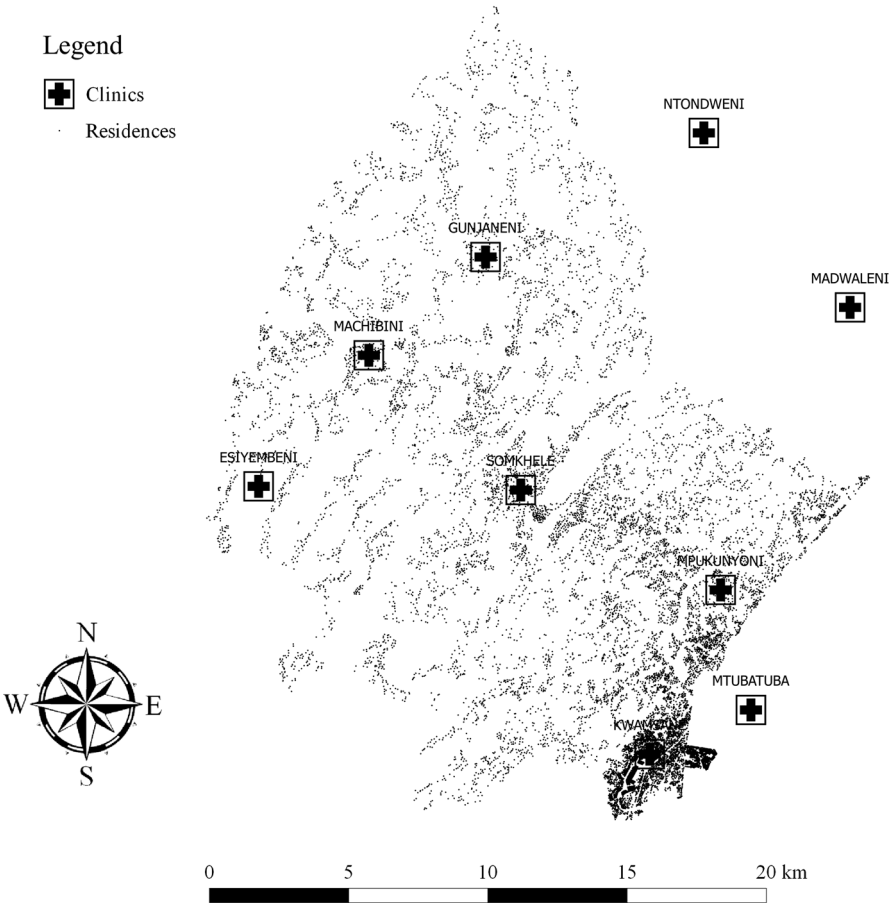

Fig. 2 Households and primary healthcare clinics. *Notes* The demographic surveillance area shown here is located in the Hlabisa sub-district in rural KwaZulu-Natal, South Africa. The area is situated in the south-east portion of the uMkhanyakude district of KwaZulu-Natal province near the town of Mtubatuba. It is bounded on the west by the Umfolozi-Hluhluwe nature reserve, on the south by the Umfolozi river, on the east by the N2 highway (except form portions where the Kwamsane township straddles the highway) and in the north by the Inyalazi river for portions of the boundary. The physical homes of local residents, locally referred to as "homesteads"

to investigate the importance of specific factors pertinent to the *nearest* facility, such as teacher quality or health staff availability. It is easy to see that scrambling will make this exercise rather difficult. As shown in Fig. 2, there are six health facilities that are located directly in the demographic surveillance area, and 15 health facilities that are located in the larger district. If household coordinates are scrambled and households are linked to the nearest health facility locations according to the scrambled household-facility distance estimate, a rather large fraction of households will be linked to an incorrect location and the correlation between the true and the actual distance will fall. This is illustrated in Table 1. With a recommended scrambling radius of 5 km, about one-third of households are linked

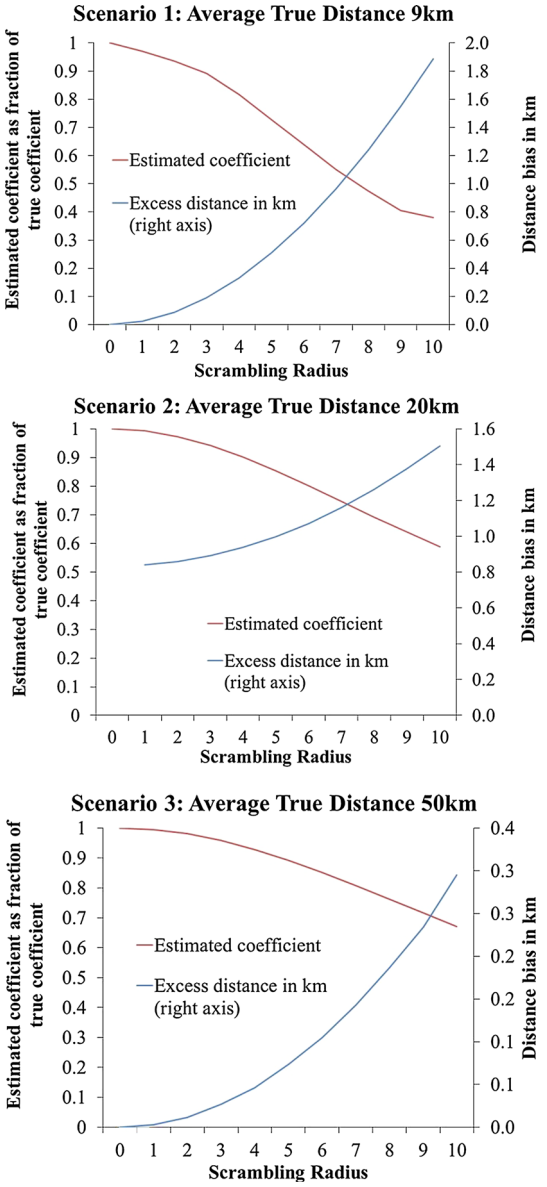

Fig. 3 Scrambling results

to the incorrect nearest facility, and the correlation between the true and the scrambled distance is <0.5.

While it is hard to generalize these biases due to their dependence on the local distribution of reference points, the added complexity of multiple reference point will in most cases substantially increase the biases generated by scrambling. Similarly, complex challenges arise if researchers want to use existing network

Table 1 Nearest primary healthcare clinics with scrambled household coordinates

Scrambling radius (km)	Correlation between true and scrambled distance	Percentage of household matched to correct facility (%)
0	1.000	100.0
1	0.965	92.1
2	0.874	85.1
3	0.748	78.5
4	0.610	72.0
5	*0.476*	*65.5*
6	0.360	59.2
7	0.264	53.3
8	0.187	48.0
9	0.124	43.3
10	0.074	39.2

5-km radius is in italics as currently recommended radius for rural areas in DHS surveys

information to measure actual (rather than Euclidean) distances traveled on roads or public transport to reach specific access points. In the case of network analysis, scrambling home locations will bias average distances and will also lead to miscoding of transport entry points, with resulting error in travel time as well as environmental risk exposure.

Summary and conclusion

In this paper, we have proved mathematically and demonstrated empirically that scrambling of GPS locations leads to a systematic overestimation of the average distance between households and other points of interest at the population level for descriptive purposes. For bi- or multivariate regression analysis, the use of scrambled GPS coordinates will lead to systematic underestimation of the true causal effects of distance. Both effects are problematic from a scientific and a policy perspective. The systematic underestimation of the true causal effect of proximity will likely undermine the perceived importance of spatial distance. This may discourage public investment that ensures geographical accessibility to essential services, such as education, health care, or transport; it may also reduce support for projects aimed at ensuring sufficient distance to nearby hazards, such as waste sites, nuclear reactors, or sources of noise pollution.

This paper is, to our knowledge, the first to fully quantify the biases resulting from scrambling. The main results presented strongly support the recommendation made in the 2007 Committee on the Human Dimensions of Global Change special report *Putting People on the Map* (2007, p. 62), which states that "[a]ltering data to mask the exact spatial locations impedes the ability of researchers to calculate accurate spatial relationships, such as distances."

While scrambling is currently common practice in major population-based surveys, such as DHS, and longitudinal surveillance systems, such as HDSS, it is only one of many approaches to mask geo-spatial data that have been proposed and used. Rather than randomly moving coordinates, one may also displace co-ordinates deterministically to a new set of locations, through displacement, scaling, or rotation. Each of these deterministic geo-masking approaches fails to preserve some important geographical information. The random perturbations generated by scrambling have previously been thought to approximately preserve all important aspects of geographical information. Unlike deterministic geo-masking, scrambling has thus been judged to be "satisfactory from a comprehensive information-preservation standpoint" (Armstrong et al. 1999). We show in this paper that scrambling does not preserve distance information. For many analytical purposes, scrambling does not appear the superior geo-masking approach it has previously been considered to be, and its routine use to mask geographical information in major population-based surveys deserves re-consideration.

Given that geo-referenced data are as important for science as they are for policymaking, and given that the protection of data confidentiality is an important dimension of ethical research, alternative approaches to handling geo-referenced data appear preferable. One commonly practiced alternative to data scrambling, which appears strongly preferable to scrambling, is the "restricted release" of data suggested by the Committee on the Human Dimensions of Global Change (2007) as well as National Research Council (2005). Such data access restrictions require an application and review process, as well as specific access rules and locations, which are comparatively costly and may limit the use of this approach to resource-rich settings.

A less costly alternative to scrambling is to provide distance calculations on request. Rather than providing researchers with geographical coordinates, distances between two points (e.g., a household and a health care facility) could be calculated by data owners and the resulting distance measures could be shared with researchers instead of the coordinates. While this would ensure privacy protection from an individual point of view, it would also provide researchers with the key variables needed for empirical analysis. For instance, to assess the effect of distance on access to antiretroviral treatment in developing countries (see e.g., Bärnighausen et al. 2014), variables such as "distance to the nearest primary healthcare clinic" or "distance to the nearest road" could be computed. Given that researchers would neither know the reference point location nor the direction from the reference point, identifying households based on these distance measures would be impossible.

An alternative promising approach is the use of software agents, which could allow researchers to analyze fully identified data without being able to physically access the underlying identifiable information (Kamel Boulos et al. 2006). Developing such system may require substantial upfront investment in order to ensure that they are user-friendly and provide sufficient data protection, but they may be the most efficient solution in the long run. In the absence of software agents, restricted release of true coordinates or true distances appears strongly preferable to the scrambling approach.

Appendix: Proof of proposition 1

It is easy to show that in the degenerate case $h = 0$ the average distance $\langle |h + v| \rangle$ must exceed $|h|$. We shall show that this behavior $\langle |h + v| \rangle > |h|$ is typical even when $|h|$ is comparable with, or considerably greater than, the typical size $|v|$ of the noise vector.

The result $\langle |h + v| \rangle > |h|$ is a direct consequence of the triangle inequality (TI). Recall that the TI asserts that for any vectors x, y, we have

$$|x| + |y| \geq |x + y|,$$

because the right-hand side is the distance from the origin to $x + y$ along a straight line, and the left-hand side is the distance from the origin to $x + y$ via x. Hence, equality $|x| + |y| = |x + y|$ occurs if and only if x is contained in the closed line segment joining the origin to $x + y$, that is, if and only if one of x and y is a non-negative multiple of the other.

To prove that $\langle |h + v| \rangle > |h|$, first note that $\langle |h + v| \rangle = \langle |h - v| \rangle$ because v and $-v$ have the same distribution. But then,

$$\begin{aligned} 2\langle |h + v| \rangle &= \langle |h + v| \rangle + \langle |h - v| \rangle \\ &= \langle |h + v| + |h - v| \rangle \\ &\geq \langle |(h + v) + (h - v)| \rangle \\ &= \langle |2h| \rangle \\ &= |2h| \\ &= 2|h|, \end{aligned}$$

using TI with $x = h + v$ and $y = h - v$ in the second step and the fact that $|2h|$ is constant in the next-to-last step. Dividing both sides of the resulting inequality by 2, we deduce $\langle |h + v| \rangle \geq |h|$ as claimed. Moreover, we can only have $\langle |h + v| \rangle = |h|$ when $h + v$ and $h - v$ satisfy the equality condition in the TI for every v, which is to say when every v is on the closed line segment joining 0 to h. To evaluate $\langle |h + v|^2 \rangle$, we begin in the same way

$$2\left\langle |h + v|^2 \right\rangle = \left\langle |h + v|^2 \right\rangle + \left\langle |h - v|^2 \right\rangle = \left\langle |h + v|^2 + |h - v|^2 \right\rangle,$$

and now apply the parallelogram identity

$$|x + y|^2 + |x - y|^2 = 2|x|^2 + 2|y|^2$$

(which can be obtained by writing each of the terms $|x \pm y|^2$ on the left-hand side as an inner product $(x \pm y, x \pm y) = (x, x) \pm 2(x, y) + (y, y)$ and noting that the cross-terms $\pm 2(x, y)$ sum to zero). Taking $x = h$ and $y = v$, we then obtain

$$2\left\langle |h + v|^2 \right\rangle = \left\langle 2|h|^2 + 2|v|^2 \right\rangle = \left\langle 2|h|^2 \right\rangle + \left\langle 2|v|^2 \right\rangle = 2|h|^2 + 2\left\langle |v|^2 \right\rangle.$$

Dividing both sides by 2, we recover the identity $\langle |h + v|^2 \rangle = |h|^2 + \langle |v|^2 \rangle$ claimed earlier.

Now recall that any real-valued random variable X satisfies $\langle X^2 \rangle \geq \langle X \rangle^2$ (the difference is the variance $\langle (X - \langle X \rangle)^2 \rangle$, which is clearly non-negative). Applying this to $X = |h + v|$, we find

$$\langle |h + v| \rangle^2 \leq \left\langle |h + v|^2 \right\rangle = |h|^2 + \left\langle |v|^2 \right\rangle \leq \left(|h| + \frac{\left\langle |v|^2 \right\rangle}{2|h|} \right)^2,$$

with strict inequality unless $\langle |v|^2 \rangle = 0$. Hence, $\langle |h + v| \rangle \leq |h| + \langle |v|^2 \rangle / (2|h|)$ as claimed.

So far, our analysis did not depend on the choice of distribution v or even on the dimension of the space. In practice, h and v are drawn from a two-dimensional space, though one may also consider one-dimensional problems as a simplified model (such as a community limited to a street or a long and narrow valley). We consider three possibilities:

1. A one-dimensional space with v drawn uniformly from the interval $[-\gamma, +\gamma]$ for some $\gamma > 0$, so h is replaced by a random number drawn uniformly from the interval $[h - \gamma, h + \gamma]$ of length 2γ centered at h.
2. A two-dimensional space with v drawn uniformly from the radius-γ circle $|v| = \gamma$ about the origin, so h is replaced by a random point at distance exactly γ from h (a random point on the circle of radius γ about h). Even if this distribution is not used in practice, it is needed for the analysis of the next case.
3. A two-dimensional space with v drawn uniformly from the radius-γ disk $|v| \leq \gamma$ about the origin, so h is replaced by a random point at distance at most γ from h (a random point in the disk of radius γ about h). In this case, our analysis requires that $\gamma \leq |h|$, but this assumption will usually be satisfied in practice.

In the one-dimensional case, the variance $\langle |v|^2 \rangle$ is given by the elementary integral

$$\frac{1}{2\gamma} \int_{-\gamma}^{+\gamma} v^2 dv = \frac{1}{2\gamma} \left[\frac{v^3}{3} \right]_{v=-\gamma}^{\gamma} = \frac{\gamma^2}{3},$$

so the added noise increases $\langle |h|^2 \rangle$ by $\gamma^2/3$. The expected distance $\langle |h + v| \rangle$ remains $|h|$ as long as $\gamma < |h|$, since then $|h + v| + |h - v| = 2h$ always. Once γ exceeds $|h|$, we distinguish two possibilities. In the first, $|h|$ still exceeds the noise magnitude $|v|$. This happens with probability $|h|/\gamma$, and then the average value of $|h + v|$ in this case is still h. The other possibility is that $|v| \geq |h|$, and then averaging $|h + v|$ with $|h - v|$ yields $|v|$. Since here $|v|$ ranges uniformly from $|h|$ to γ, its average value is $(\gamma + |h|)/2$. Combining the $|v| < |h|$ and $|v| \geq |h|$ averages, weighted by their respective probabilities, we obtain

$$\frac{|h|}{\gamma} |h| + \left(1 - \frac{|h|}{\gamma} \right) \frac{\gamma + |h|}{2} = \frac{\gamma^2 + |h|^2}{2\gamma} = h + \frac{(\gamma - |h|)^2}{(2\gamma)}.$$

Thus, replacing h by $h + v$ increases the expected distance by $\frac{(\gamma-|h|)^2}{(2\gamma)}$.

In the second scenario, $|v| = \gamma$ always, so $\langle |v|^2 \rangle = \gamma^2$ and $\langle |h + v|^2 \rangle = |h|^2 + \gamma^2$. To compute $\langle |h + v| \rangle$, let $\theta \in [0, 2\pi)$ be the oriented angle from h to v. Then, θ is uniformly distributed in $[0, 2\pi)$ and $|h + v| = \sqrt{|h|^2+2\gamma|h|\cos\theta+\gamma^2}$ by the Law of Cosines [or by expanding the inner product of $|h + v|^2 = (h + v, h + v) = (h, h) + 2(h, v) + (v, v)$]. Thus,

$$\langle |h + v| \rangle = \frac{1}{2\pi} \int_0^{2\pi} \sqrt{|h|^2+2\gamma|h|\cos\theta+\gamma^2}\, d\theta.$$

This integral is no longer elementary, except in the special case where $\gamma = 0$, $h = 0$, or $\gamma = |h|$. [If $\gamma = 0$ or $h = 0$ then $\langle |h + v| \rangle = |h|$ or γ, respectively; if $\gamma = |h|$, then the identity $2 + 2\cos\theta = 4\cos^2(\theta/2)$ simplifies the integral to $\int_0^{2\pi} 2|h||\cos(\theta/2)|d\theta = 8|h|$, whence $\langle |h + v| \rangle = (4/\pi)|h|$]. Assume, then, that 0, γ, and $|h|$ are distinct. Then, we may assume $\gamma < |h|$ because our formula for $\langle |h + v| \rangle$ does not change if we switch γ with $|h|$. Then, our integral can be evaluated in terms of a complete elliptical integral of the second kind[2]:

$$\int_0^{2\pi} \sqrt{|h|^2+2\gamma|h|\cos\theta+\gamma^2}\, d\theta = 4(|h| + \gamma)\mathbf{E}'\left(\frac{|h| - \gamma}{|h| + \gamma}\right).$$

It would take substantial work to recover the behavior of $\langle |h + v| \rangle$ from this rather exotic formula. We thus work directly with the integral, expanding it as a power series in γ that converges in the interval $|\gamma| < |h|$.

It will be convenient to regard h and v as complex numbers in the usual way. Then, $h + v = h(1 + re^{i\theta})$, where $r = \gamma/h < 1$. We then have

$$|h + v| = |h|\left|1 + re^{i\theta}\right| = |h|\left(\left(1 + re^{i\theta}\right)\overline{\left(1 + re^{i\theta}\right)}\right)^{\frac{1}{2}},$$

and since the complex conjugate of $1 + re^{i\theta}$ is $1 + re^{-i\theta}$, this gives

$$|h + v| = |h|\left(\left(1 + re^{i\theta}\right)\left(1 + re^{-i\theta}\right)\right)^{\frac{1}{2}} = |h|\sqrt{1 + re^{i\theta}}\sqrt{1 + re^{-i\theta}}.$$

We expand each of the factors $\sqrt{1 + re^{\pm i\theta}}$ using the binomial series

$$\sqrt{1 + z} = (1 + z)^{\frac{1}{2}} = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \ldots,$$

valid and absolutely convergent for all complex z such that $|z| \leq 1$, where

[2] See for instance formulas 8.111#3 and 8.112#2 of: I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products* (tr. and ed. Alan Jeffrey), New York: Academic Press 1980. The power series in γ that we obtain is probably known too, but easier to derive than to locate in the literature.

$$a_0 = 1, a_1 = \frac{1}{2}, a_2 - \frac{1}{8}, a_3 = \frac{1}{16}, a_4 = -\frac{5}{128}, a_5 = \frac{7}{256},$$

and in general the coefficient a_m is $\left(\frac{1}{2}\right)\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\ldots\left(-m+\frac{3}{2}\right)/m!$. Thus, $\sqrt{1+re^{i\theta}}\sqrt{1+re^{-i\theta}}$ is the sum of the terms $a_m a_n r^{m+n} e^{i(m-n)\theta}$ over all pairs (m, n) of whole numbers. The integral of such a term over $0 \le \theta \le 2\pi$ is $2\pi a_m a_n$ if $m = n$ and zero otherwise. Summing over m, n, we find that $4(|h| + \gamma)\mathbf{E}'\left(\frac{|h-\gamma|}{h+\gamma}\right)$ is the sum of the terms $2\pi|h|a_n^2 r^{2n}$ over $n = 0, 1, 2, 3,\ldots$, and thus that

$$\langle |h+v| \rangle = |h|\left(a_0^2 + a_1^2 r^2 + a_2^2 r^4 + a_3^2 r^6 + \cdots\right)$$

$$= |h| + \frac{1}{4}\frac{\gamma^2}{|h|} + \frac{1}{64}\frac{\gamma^4}{|h|^3} + \frac{1}{256}\frac{\gamma^6}{|h|^5} + \frac{25}{16384}\frac{\gamma^8}{|h|^7} + \cdots$$

We note in passing that the special case $\gamma = |h|$ (that is, $r = 1$) yields the amusing formula

$$\frac{4}{\pi} = \sum_{n=0}^{\infty} a_n^2 = 1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{1}{16}\right)^2 + \left(\frac{5}{128}\right)^2 + \left(\frac{7}{256}\right)^2 \cdots$$

In the more general and final scenario 3, v is drawn uniformly from the radius-γ disk $|v| \le \gamma$ about the origin. We integrate over this circle using polar coordinates, again using for θ the oriented angle from h to v. We then find

$$\left\langle |v|^2 \right\rangle = \frac{1}{\pi\gamma^2}\int_{\rho=0}^{\gamma} \rho \int_{\theta=0}^{2\pi} \rho^2 d\theta d\rho = \frac{2\pi}{\pi\gamma^2}\int_{\rho=0}^{\gamma} \rho^3 d\rho = \frac{2}{\gamma^2}\left[\frac{\rho^4}{4}\right]_{\rho=0}^{\gamma} = \frac{\gamma^2}{2},$$

and thus $\left\langle |h+v|^2 \right\rangle = |h|^2 + \frac{1}{2}\gamma^2$. For the average distance, we write

$$\langle |h+v| \rangle = \frac{1}{\pi\gamma^2}\int_{\rho=0}^{\gamma} \rho \int_{\theta=0}^{2\pi} \sqrt{|h|^2 + 2\gamma|h|\cos\theta + \gamma^2}\, d\theta\, d\rho.$$

Again the integral is not elementary. As long as $\gamma < |h|$, we have $\rho \le |h|$ for all ρ in $[0, \gamma]$, so we can use our power series for the integral over θ and integrate each term $2\pi|h|a_n^2(\rho/h)^{2n}$ (with $n = 0, 1, 2, 3,\ldots$), obtaining

$$\frac{1}{\pi\gamma^2} 2\pi|h|^{1-2n} a_n^2 \int_{\rho=0}^{\gamma} \rho^{2n+1} d\rho = \frac{2|h|^{1-2n} a_n^2}{\gamma^2}\left[\frac{\rho^{2n+2}}{2n+2}\right]_{\rho=0}^{\gamma} = |h|\frac{a_n^2}{n+1} r^{2n},$$

where $r = \gamma/|h|$ as before. Therefore, in this case, we obtain the power series expansion

$$\langle |h+v| \rangle = |h|\left(a_0^2 + \frac{a_1^2}{2} r^2 + \frac{a_1^2}{3} r^4 + \frac{a_1^2}{4} r^6 + \cdots\right)$$

$$= |h| + \left(\frac{1}{8} \frac{\gamma^2}{|h|} + \frac{1}{192} \frac{\gamma^4}{|h|^3} + \frac{1}{1024} \frac{\gamma^6}{|h|^5} + \frac{5}{16384} \frac{\gamma^8}{|h|^7} + \cdots \right).$$

References

Arcury, T. A., Gesler, W. M., Preisser, J. S., Sherman, J., Spencer, J., & Perin, J. (2005). The effects of geography and spatial behavior on health care utilization among the residents of a rural region. *Health Services Research, 40*(1), 135–155. doi:10.1111/j.1475-6773.2005.00346.x.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine, 18*(5), 497–525.

Bärnighausen, T., Tanser, F., Herbst, K., Mutevedzi, T., Mossong, J., & Newell, M. (2014). Structural barriers to antiretroviral treatment: A study using population-based CD4 count and linked antiretroviral treatment programme data. *Lancet, 382*, S5.

Borsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., et al. (2013). Data resource profile: The survey of health, ageing and retirement in Europe (SHARE). *International Journal of Epidemiology, 42*(4), 992–1001. doi:10.1093/ije/dyt088.

Center for Human Resource Research. (1997). *The national longitudinal surveys NLSY79 user guide*. Columbus, OH: Ohio State University.

Committee on the Human Dimensions of Global Change. (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data*. Washington, DC: The National Academic Press.

Cooke, G. S., Tanser, F. C., Bärnighausen, T., & Newell, M. L. (2010). Population uptake of antiretroviral treatment through primary care in rural South Africa. *BMC Public Health, 10*, 585. doi:10.1186/1471-2458-10-585.

Golden, M. L., Downs, R. R., & Davis-Packard, K. (2005). *Confidentiality issues and policies related to the utilization and dissemination of geospatial data for public health applications*. New York: The Socioeconomic Data and Applications Center (SEDAC), Center for International Earth Science Information Network (CIESIN), Columbia University.

Hyman, S. E. (2000). The needs for database research and for privacy collide. *American Journal of Psychiatry, 157*(11), 1723–1724.

ICF International. (2012). *Demographic and health survey—Sampling and household listing manual MEASURE DHS*. Calverton, Maryland, USA

Kamel Boulos, M. N., Cai, Q., Padget, J. A., & Rushton, G. (2006). Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics, 39*(2), 160–170. doi:10.1016/j.jbi.2005.06.003.

Kamel Boulos, M. N., Curtis, A. J., & Abdelmalik, P. (2009). Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics, 8*, 46. doi:10.1186/1476-072X-8-46.

Kaplan, E. D., & Hegarty, C. J. (2006). *Understanding GPS: Principles and applications* (2nd ed.). Boston: Artech House.

Kyei, N. N., Campbell, O. M., & Gabrysch, S. (2012). The influence of distance and level of service provision on antenatal care use in rural Zambia. *PLoS ONE, 7*(10), e46475. doi:10.1371/journal.pone.0046475.

Linardakis, M., Smpokos, E., Papadaki, A., Komninos, I. D., Tzanakis, N., & Philalithis, A. (2013). Prevalence of multiple behavioral risk factors for chronic diseases in adults aged 50+ , from eleven European countries—The SHARE study (2004). *Preventive Medicine, 57*(3), 168–172. doi:10.1016/j.ypmed.2013.05.008.

Lohela, T. J., Campbell, O. M., & Gabrysch, S. (2012). Distance to care, facility delivery and early neonatal mortality in Malawi and Zambia. *PLoS ONE, 7*(12), e52110. doi:10.1371/journal.pone.0052110.

National Archives and Records Administration. (1996). *U.S. global positioning system policy*. Washington, DC: U.S. Government.

National Research Council. (2005). *Expanding access to research data: Reconciling risks and opportunities*. Washington, DC: The National Academies Press.

O'Brien, D. G., & Yasnoff, W. A. (1999). Privacy, confidentiality, and security in information systems of state health agencies. *American Journal of Preventive Medicine, 16*(4), 351–358.

Onsrud, H. J., Johnson, J. P., & Lopez, X. (1994). Protecting personal privacy in using geographic information systems. *Photogrammetric Engineering and Remote Sensing, 60*(9), 1083–1095.

Schwieger, V. (2003). Using handheld GPS receivers for precise positions. *FIG Regional Conference Paper.*

Seiber, E. E., & Bertrand, J. T. (2002). Access as a factor in differential contraceptive use between Mayans and ladinos in Guatemala. *Health Policy and Planning, 17*(2), 167–177.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72–101.

Tanser, F., Bärnighausen, T., Cooke, G. S., & Newell, M. L. (2009). Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. *International Journal of Epidemiology, 38*(4), 1008–1016. doi:10.1093/ije/dyp148.

Tanser, F., Hosegood, V., Barnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., et al. (2008). Cohort profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *International Journal of Epidemiology, 37*(5), 956–962. doi:10.1093/ije/dym211.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach* (2nd ed.). South-Western: Thomson.