Modeling distances between humans using Taylor's law and geometric probability

Joël Cohen, Daniel Courgeau

Modeling distances between humans using Taylor's law and geometric probability

Joel E. Cohen
Laboratory of Populations, The Rockefeller University and Columbia University, New York, USA

Daniel Courgeau
Institut National d'Etudes Démographiques, Paris, France

Abstract

Taylor's law states that variance of the distribution of distance between two randomly chosen individuals is a power function of the mean distance. It applies to the distances between two randomly chosen points in various geometric shapes, subject to a few conditions. In Réunion Island and France, at some spatial scales, the empirical frequency distributions of inter-individual distances are predicted accurately by the theoretical frequency distributions of inter-point distances in models of geometric probability under a uniform distribution of points. When these models fail to predict the empirical frequency distributions of inter-individual distances, they provide baselines against which to highlight the spatial distribution of population concentrations.

Keywords: geometric probability; Taylor's law; spatial distribution; population density; urban hierarchies; internal migration

The published paper may be consulted on the Taylor & Francis website:

http://www.tandfonline.com/doi/full/10.1080/08898480.2017.1289049

The order of authorship is alphabetical. Both authors contributed equally.
Address correspondence to Daniel Courgeau, 114-2 chemin du défends, Mougins, 06250, France. Email: daniel.courgeau@wanadoo.fr

Introduction

In animal ecology, Taylor (1961, 1986) compared aggregation from the distribution of population densities across sub-groups of a given population. In demography, Courgeau (1970, 1973) and Bell et al. (2015) compared internal migrations from the distribution of distances between randomly chosen individuals in human populations from different countries. Here we relate these two perspectives on the spatial distribution of populations. We link empirical data from human populations to models based on geometric probability.

Taylor's (1961) law (also known as Taylor's power law) asserts that the logarithm of the variance Var(X) of the density X of a set of comparable groups is a linear function of the logarithm of the mean density $E(X)$, or equivalently that there exist constants $a > 0$ and b such that

$$\mathrm{Var}\left(X\right) = a\ E\left(X\right)^{b}. \tag{1}$$

Taylor's law has been verified, using the sample mean and the sample variance as estimates of the population mean and variance, in cosmology, with the emission spectra of x-ray binary systems and active galactic nuclei (Uttley and McHardy, 2001); in ecology, with population densities of bacteria (Ramsayer et al., 2012) or plants, insects, and animals (Taylor et al., 1978); and in social sciences, with the spatial distribution of human population densities (Cohen et al., 2013).

Social groups are organized according to rules generating spatial structures. Examples include urbanization for humans and colonies for termitaries and anthills. Different regions, including the different subgroups considered in Taylor's law, may not be independent of one another in the presence of migration. The distribution of the distances between pairs of randomly chosen individuals is another way of describing the spatial distribution of

populations. It has been modeled as the distribution of the distances between two randomly chosen points in various geometric regions, using geometric probability.

Geometric probability originated with Buffon (1733), who studied the spatial distribution of different kinds of randomly distributed objects. Crofton (1885) calculated the probability that a figure formed by *n* points randomly distributed on a given surface possesses a specific property independently of any overall translation or rotation. One such property concerns the distance between a pair of randomly selected points. Barton et al. (1963) in chromosome analysis, Kuiper and Paelinck (1982) in geography, Moltchanov (2012) in network analysis; Courgeau (1970), Courgeau and Baccaïni (1989), Rogerson (1990) in demography, and Parry and Fischbach (2000) in physics have addressed such questions. However, these models use uniform distributions and only approximate non-uniform distributions with a finite number of points. Spatial point process models use geometric probability to describe the arrangement and interactions of objects unevenly or randomly distributed on a plane or in a space (Boutin and Kemper, 2004; Illian et al., 2008).

We show that the distribution of distances between two random individuals obeys Taylor's law with exponent $b = 2$ in a Euclidean space of any finite dimension, for all regions of the same shape as a given bounded region of any shape (section 1). The space may be continuous or discrete and the total number of points may be finite or infinite, subject to a few conditions. The case $b = 2$ has special significance. When $b = 2$, the coefficient of variation, which is the ratio of the standard deviation $(\mathrm{Var}(X))^{1/2}$ of the density X to the mean density $E(X)$, equals $a^{1/2}$ for all values of $E(X)$. For nonnegative random variables such as population density X, the coefficient of variation equals the reciprocal of the signal-to-noise ratio, which is the mean divided by the standard deviation. When $b = 2$, the signal-to-noise

ratio equals $a^{-1/2}$ for all values of $E(X)$. Conversely, when the coefficient of variation is constant, then Taylor's law holds with $b = 2$.

We shall review and extend results of geometric probability for different territorial shapes, including an annulus (section 2). We confirm that the distribution of distances between individuals obeys Taylor's law with exponent $b = 2$ in these examples. In section 3, we consider a finite space and a finite number of individuals to examine the effect of spatial variation in population density and how these results are modified by territorial and social constraints. We give simulated and empirical examples (sections 2 and 3).

1. Application of Taylor's law with geometric probability

The intuition behind the theorem in section 1.2 is that for any family of similar shapes, both the mean distance between two randomly chosen points and the standard deviation of the distance between two randomly chosen points scale linearly with rescaling (e.g., by changing the radius of a circle or the side of a regular polygon, and similarly for any shape), and therefore the variance scales as the second power of the mean. This intuition applies to all the cases in section 2.

1.1 Definitions and preliminary results

In plane geometry, two shapes are defined to be "similar" if one can be perfectly transformed to another by rescaling, rotation, reflection (mirror image), and translation. For example, all equilateral triangles are similar, all circles are similar, and all squares are similar, but no equilateral triangle is similar to a square or to a right triangle.

$\mathbb{R}$ is the real line, and N any fixed positive integer. $\mathbb{R}^N$ is the N-dimensional space of real vectors $x = (x_1, ..., x_N)$. For any two points x and y in $\mathbb{R}^N$, the Euclidean distance between them is

$$r(x,y)=\left(\sum_{i=1}^{N}(x_i-y_i)^2\right)^{\frac{1}{2}}. \tag{2}$$

The Euclidean N-space ($\mathbb{R}^N$, $r(.,.)$) is the set $\mathbb{R}^N$ of N-dimensional real points together with the metric $r(.,.)$, which gives the Euclidean distance $r(x, y)$ between a pair of points x and y in $\mathbb{R}^N$.

S is any fixed set of points in $\mathbb{R}^N$. It may be continuous or discrete, and the total number of points in S may be finite or infinite. For any two points x and y in S, $r(x, y)$ is not necessarily bounded, though by definition of $\mathbb{R}^N$, $r(x, y) = \infty$ is excluded. For convenience, we take S as continuous. $\mathbb{R}_+$ is the set of positive finite real numbers $(0, +\infty)$. For any c in $\mathbb{R}_+$, define the rescaling of S by c as the set of points $cS \equiv \{cx \,|\, x \in S\}$. Define the family of S, $\mathcal{S}(S)$, to comprise all sets of points in $\mathbb{R}^N$ obtained from S by rescaling S by c, that is, $\{cS \mid c \in \mathbb{R}_+\}$, and all translations and all rotations of all rescalings of S.

For example, if $N = 1$ and $S = [0, 1]$, then the family $\mathcal{S}(S)$ consists of all closed intervals of the real line $\mathbb{R}$. If $N = 2$ and S is the unit disk, then $\mathcal{S}(S)$ is the family of all closed disks of any radius centered anywhere in the plane.

Consider a probability density function p_S on the points of S:

$$\text{for any } x \text{ in } S,\ 0 \le p_S(x) \text{ and } \int_{x\in S} p_S(x)\,dx = 1. \tag{3}$$

If S is a set in $\mathbb{R}^N$, then the infinitesimal volume element is

$$\prod_{i=1}^{N} dx_i\,. \tag{4}$$

Define the random distance X_S to be the random variable constructed by picking x in S with probability $p_S(x)\,dx$, independently picking y in S with probability $p_S(y)\,dy$, and computing the Euclidean distance $r(x, y)$ between the chosen points. Define the mean and the standard deviation of the random distance X_S on S as

$$\mu(S)=E(X_S)=\int_{x,\,y\in S} r(x,y)\,p_S(x)\,p_S(y)\,dx\,dy, \tag{5}$$

$$\sigma(S)=\left(E\left((X_S-\mu(S))^2\right)\right)^{\frac{1}{2}}=\left(\int_{x,\,y\in S}(r(x,y)-\mu(S))^2\;p_S(x)\;p_S(y)\,dx\,dy\right)^{\frac{1}{2}}. \tag{6}$$

We assume that $0 < \mu(S)$, thus excluding S consisting of a single point, because $r(x,x) = 0$. We also assume that $0 < \sigma(S)$, thus excluding S consisting of exactly two points or the vertices of an equilateral triangle. We finally assume that $\mu(S) < \infty$, and that $\sigma(S) < \infty$. The last two conditions hold true if S is bounded (there exists k in $\mathbb{R}_+$ such that, for all x, y in S, $r(x,y) \le k$) but may also hold true if S is not bounded, depending on the probability density function $p(.)$.

Under these assumptions, we define the variance of X_S on S as the square of its standard deviation: $\mathrm{Var}(S) = \sigma^2(S)$. The coefficient of variation of the random distance X_S on S is

$$\mathrm{CV}(S) = \sigma(S)/\mu(S). \tag{7}$$

Then $0 < \mathrm{CV}(S) < \infty$.

Lemma 1

Consider x and y two points in $\mathbb{R}^N$. Then for any c in $\mathbb{R}_+$, $r(cx,cy) = cr(x,y)$. That is, rescaling any two points x, y by the factor c rescales the Euclidean distance $r(x, y)$ by the same factor c. Rotating around a point and translating x and y by the same vector have no effect on the distance between them.

Proof.

$$r(cx,cy) = \left(\sum_{i=1}^{N} (cx_i - cy_i)^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{N} c^2 (x_i - y_i)^2 \right)^{\frac{1}{2}} = c\, r(x,y). \tag{8}$$

Lemma 2

If cS is a rescaling of S by c in $\mathbb{R}_+$, define the probability density function p_{cS} on the points of cS by $p_{cS}(cx) = p_S(x)/c^N$ for all points cx in cS, or equivalently for all points x in S. Then p_{cS} is a probability density function on cS.

Proof.

By construction, $p_{cS}(cx) \ge 0$. Because the infinitesimal of the volume element $d(cx)$ in cS is the product of the infinitesimals of each dimension,

$$d(cx) = \prod_{i=1}^{N} d(cx_i) = \prod_{i=1}^{N} (c\,dx_i) = c^N\,dx, \tag{9}$$

we have

$$\int_{y=cx\in cS} p_{cS}(y)\,dy = \int_{x\in S} \left(p_S(x)/c^N\right) d(cx) = \int_{x\in S} \left(p_S(x)/c^N\right) c^N d(x) = \int_{x\in S} p_S(x)\,dx = 1. \tag{10}$$

1.2 Theorem and theoretical example

S is a set of at least three points in the N-dimensional Euclidean space $(\mathbb{R}^N, r(.,.))$ such that not all pairs of points are equidistant. $\mathcal{S}(S)$ is the infinite family of sets of points in $\mathbb{R}^N$ including all rescalings of S, namely, $\{\, cS \,|\, c \in \mathbb{R}_+ \,\}$, all translations, all reflections, and all rotations of all rescalings, translations, and reflections of S. With a probability density function p_S on the points of S and the probability density function p_{cS} on the points of cS defined by $p_{cS}(cx) = p_S(x)/c^N$ for all points cx in cS, for any set s in $\mathcal{S}(S)$, $\mu(s)$ is the mean and $\sigma(s)$ the standard deviation of the distance between two randomly chosen points. Assume $0 < \mu(S) < \infty$ and $0 < \sigma(S) < \infty$. Then

$$\mu(s) = c\mu(S),\ \sigma(s) = c\sigma(S),\ \mathrm{CV}(s) = \mathrm{CV}(S). \tag{11}$$

The last equality in (11) holds independently of c, and the variance of the random distance on s satisfies

$$\mathrm{Var}(s) = (\mathrm{CV}(S))^2 \times (\mu(s))^2. \tag{12}$$

This is Taylor's law with $a = (\mathrm{CV}(S))^2$ and exponent $b = 2$.

Proof. Because rotations, reflections, and translations have no effect on the distance $X(cS)$, we need to examine only the effects of rescaling S by c in $\mathbb{R}_+$. By definition of the mean $\mu(.)$, using $p_{cS}(cx) = p_S(x)/c^N$ for all points cx in cS or equivalently for all points x in S

$$\begin{aligned}
\mu(s) &= E(X_{cS}) = \int_{u,v\in cS} r(u,v)\,p_{cS}(u)\,p_{cS}(v)\,du\,dv \\
&= \int_{x,y\in S} r(cx,cy)\,p_{cS}(cx)\,p_{cS}(cy)\,c^N dx\,c^N\,dy \\
&= \int_{x,y\in S} cr(x,y)\left(\frac{p_S(x)}{c^N}\right)\left(\frac{p_S(y)}{c^N}\right) c^N dx\,c^N\,dy \\
&= c\int_{x,y\in S} r(x,y)\,p_S(x)\,p_S(y)\,dx\,dy = c\mu(S).
\end{aligned} \tag{13}$$

The proof that $\sigma(s) = c\,\sigma(S)$ follows the same steps, starting from the definition of the standard deviation. Dividing the equation $\sigma(s) = c\,\sigma(S)$ by the equation $\mu(s) = c\,\mu(S)$ gives

$$\mathrm{CV}(s) = \mathrm{CV}(S), \tag{14}$$

which implies $\sigma(s) = \mathrm{CV(S)}\,\mu(s)$. Squaring both sides of equation (14) gives $\mathrm{Var}(s) = \sigma^2(s) = (\mathrm{CV(s)})^2\,(\mu(s))^2$, which is Taylor's law with exponent $b = 2$. □

Nothing in this theorem requires that the shape S be topologically connected.

The space can be extended from real to complex with an appropriate Euclidean distance, the Minkowski 2-norm,

$$|x - y|_2 = \left(\sum_{i=1}^{N} \left(|x_i - y_i| \right)^2 \right)^{\frac{1}{2}}. \tag{15}$$

The theorem extends to all Minkowski p-norms

$$|x - y|_p = \left(\sum_{i=1}^{p} \left(|x_i - y_i| \right)^p \right)^{\frac{1}{p}}, \tag{16}$$

for any $1 \leq p < \infty$ (Steele, 2004: 140), or to any homogeneous metric $r(x, y)$ such that $r(cx, cy) = c\,r(x, y)$.

As an example, consider the whole plane $\mathbb{R}^2$, and the normal distribution $N(\mu, c^2)$ on $\mathbb{R}^2$ with mean μ and standard deviation c. μ has no effect on the difference between two points, so the distribution of the distance $X(\mu, c^2)$ between two points randomly selected from $N(\mu, c^2)$ is the same for any μ. Increasing or decreasing c leaves the geometric region (which here is the whole plane) unchanged but spreads or contracts the domain of the probability density function over the plane. This example satisfies the assumptions of the theorem. Therefore $X(\mu, c^2)$, the random distance between a randomly selected pair of points, satisfies Taylor's law with $b = 2$ for all $c > 0$, translations, reflections, and rotations.

Because the difference between two independent normal variables, each of which is identically distributed as $N(\mu, c^2)$, has the distribution $N(0, 2c^2)$, the absolute value of that difference has the so-called "half-normal" distribution with $E(X(\mu, c^2)) = 2c/\pi^{1/2} \approx 1.1284\,c$

and $\mathrm{Var}\left(X\left(\mu, c^2\right)\right) = 2c^2\left(1 - 2/\pi\right)$. Thus Taylor's law $\mathrm{Var}\left(X\left(\mu, c^2\right)\right) = a\, E\left(X\left(\mu, c^2\right)\right)^b$ holds true with $a = \pi/2 - 1$, $b = 2$ for all $c > 0$.This example can be extended to a bivariate normal distribution on $\mathbb{R}^2$ with or without correlation between the x- and the y-axes.

2. Population uniformly distributed across a territory

We review some distributions of distances between two randomly chosen points and calculate their expected value and variance to show Taylor's law with $b = 2$, now assuming that the population is uniformly distributed across the territory.

2.1 Segment of the real line

Consider a linear territory of length $R > 0$. The distance X between two points chosen at random on the line has the probability density function f (Borel, 1924):

$$f(r) = \frac{2}{R}\left(1 - \frac{r}{R}\right),\ 0 \le r \le R\,. \tag{17}$$

Integrating with respect to r yields the cumulative distribution function F of the random variable X:

$$F(r) = \frac{2}{R}\left(r - \frac{r^2}{2R}\right),\ 0 \le r \le R\,. \tag{18}$$

The mean distance between individuals is

$$E(X) = \frac{R}{3}, \tag{19}$$

and its variance

$$\mathrm{Var}(X) = \frac{R^2}{18}. \tag{20}$$

This distribution obeys Taylor's law with exponent $b = 2$, because $\mathrm{Var}(X) = \frac{1}{2}\left(E(X)\right)^2$.

2.2 Disk and sphere

On a two-dimensional disk of radius R, the distance r between two points chosen at random has the probability density function f (Borel, 1924; Garwood, 1947; Luu Mau Thanh, 1962; Barton et al., 1963; Moltchanov, 2012):

$$f(r) = \frac{4r}{\pi R^2}\left(\operatorname{Arccos}\left(\frac{r}{2R}\right) - \frac{r}{2R}\left(1 - \frac{r^2}{4R^2}\right)^{1/2}\right). \tag{21}$$

The cumulative distribution function is

$$F(r) = 1 + \frac{2}{\pi}\left(\frac{r^2}{R^2} - 1\right)\operatorname{Arccos}\left(\frac{r}{2R}\right) - \frac{r}{\pi R}\left(1 + \frac{r^2}{2R^2}\right)\left(1 - \frac{r^2}{4R^2}\right)^{1/2}. \tag{22}$$

The mean distance between individuals is

$$E(X) = \frac{128R}{45\pi} \approx 0.9054R, \tag{23}$$

and the variance is

$$\operatorname{Var}(X) = R^2 - \left(\frac{128R}{45\pi}\right)^2 \approx 0.0934R^2. \tag{24}$$

This distribution obeys Taylor's law, with $b = 2$.

All moments of order p may be written (Tu and Fischbach, 2002):

$$\mu_p = 2^{p+2}\left(\frac{2}{p+2}\right)\frac{B\left(\frac{3}{2}, \frac{3+p}{2}\right)}{B\left(\frac{3}{2}, \frac{1}{2}\right)}R^p, \tag{25}$$

where $B(p, q)$ is the beta function, leading, for $a_p > 0$, to the relationship

$$\mu_p = a_p(\mu_1)^p, \tag{26}$$

which is Taylor's law for all moments (Giometto et al., 2015).

For a three-dimensional sphere, we have (Borel, 1924):

$$f(r)=\frac{3}{R^3}\left(r^2-\frac{3r^3}{4R}+\frac{r^5}{16R^3}\right). \tag{27}$$

The cumulative distribution function is

$$F(r)=\frac{3}{R^3}\left(\frac{r^3}{3}-\frac{3r^4}{16R}+\frac{r^6}{96R^3}\right). \tag{28}$$

The mean distance between individuals is

$$E(X)=\frac{36}{35}R\approx 1.0286\,R, \tag{29}$$

and its variance is

$$\mathrm{Var}(X)=1.2R^2-\left(\frac{36}{35}R\right)^2\approx 0.142R^2. \tag{30}$$

This distribution obeys Taylor's law with exponent 2.

2.3 Square

In a square of side R (Borel, 1924; Garwood, 1947; Luu Mau Thanh, 1962), the distribution of the distance r between two randomly selected points has probability density function:

$$f(r)=\frac{2r}{R^4}\left(\pi R^2-4Rr+r^2\right),\qquad 0\le r\le R, \tag{31}$$

$$f(r)=\frac{2r}{R^4}\left(2R^2\left(\mathrm{Arcsin}\frac{R}{r}-\mathrm{Arccos}\frac{R}{r}-1\right)+4R\left(r^2-R^2\right)^{1/2}-r^2\right),\qquad R<r\le R\sqrt{2}. \tag{32}$$

The mean distance is

$$E(X)=\frac{R}{15}\left(2+\sqrt{2}+5\ln\left(1+\sqrt{2}\right)\right)\approx 0.5214\,R \tag{33}$$

and its variance is

$$\mathrm{Var}(X) = \frac{R^2}{225}\left(69 - 4\sqrt{2} - 10\left(2+\sqrt{2}\right)\ln\left(1+\sqrt{2}\right) - 25\ln^2\left(1+\sqrt{2}\right)\right) \approx 0.0615R^2. \qquad (34)$$

This distribution obeys Taylor's law with $b = 2$.

To show how the different shapes affect the probability density functions, Figure 1 presents the curves for a circle and a square with the same mean distance 0.5. For the circle, the diameter equals 1.10448; for the square, the side equals 0.95895 and the maximum distance is 1.3562.

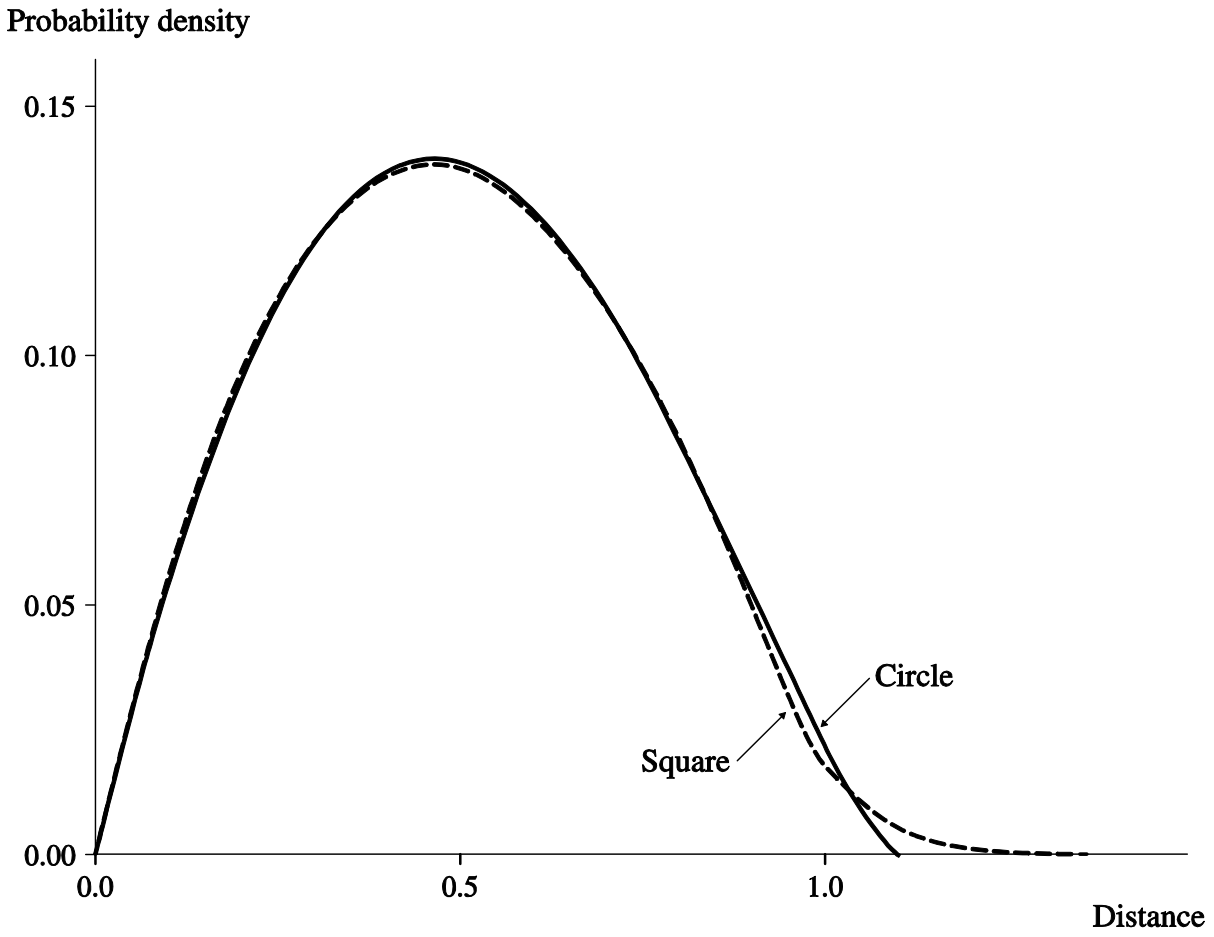
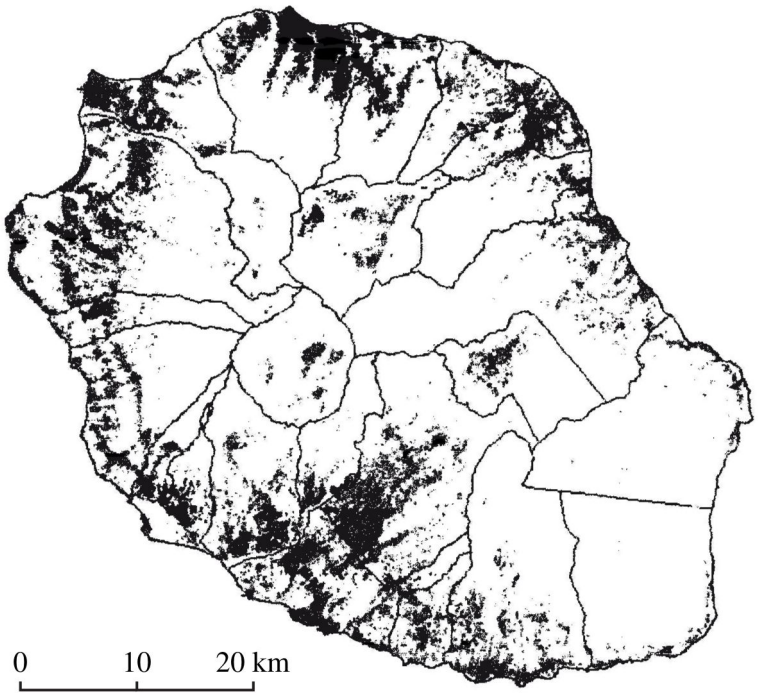


Figure 1. Distribution of distances between two random points from a circle (solid line) and a square (dashed line) with the same mean distance of 0.5.

Only the tail of the distribution differs. If two points are randomly selected with a uniform distribution from a country with boundaries that are more complicated than those of a circle or a square, only the tails of the distribution of distances between those two points are likely to be substantially affected. The core of the distribution is likely to remain unchanged.

2.4 Annulus

Consider an annulus (the area between two concentric circles) with an inner radius of R_1 and an outer radius of R_2. To our knowledge this case has never been studied until now. The Réunion Island offers a real-life approximation. Its central region—occupied by high mountains and an active volcano—is uninhabitable and its territory has a roughly circular shape. The humanly habitable zone of the island has the shape of an annulus (Figure 2).



Source: Bénard, 2012.

Map 1. Inhabited areas in Réunion Island

We seek the distribution of the straight-line distance between two random points *P* and *Q* drawn independently of one another in the annulus. An alternative to the linear distance between two points is a curvilinear path running within the annulus, but calculations are more complicated. The angle formed by $\overrightarrow{PQ}$ and a direction fixed at the beginning of the calculations is uniformly distributed in the interval $[0, 2\pi]$. If the angle lies in the interval $[\omega, \omega + d\omega)$ and if the distance |PQ| lies in the interval $[r, r + dr)$, then *P*, whose position is constrained by *r*, must lie in the set of areas *S* common to the initial annulus and to the annulus obtained by the translation by vector $\overrightarrow{QP}$ (Figure 2). For a given position P, Q is located in the area element $r\,dr\,d\omega$, and the different angles are given in Figure 3.

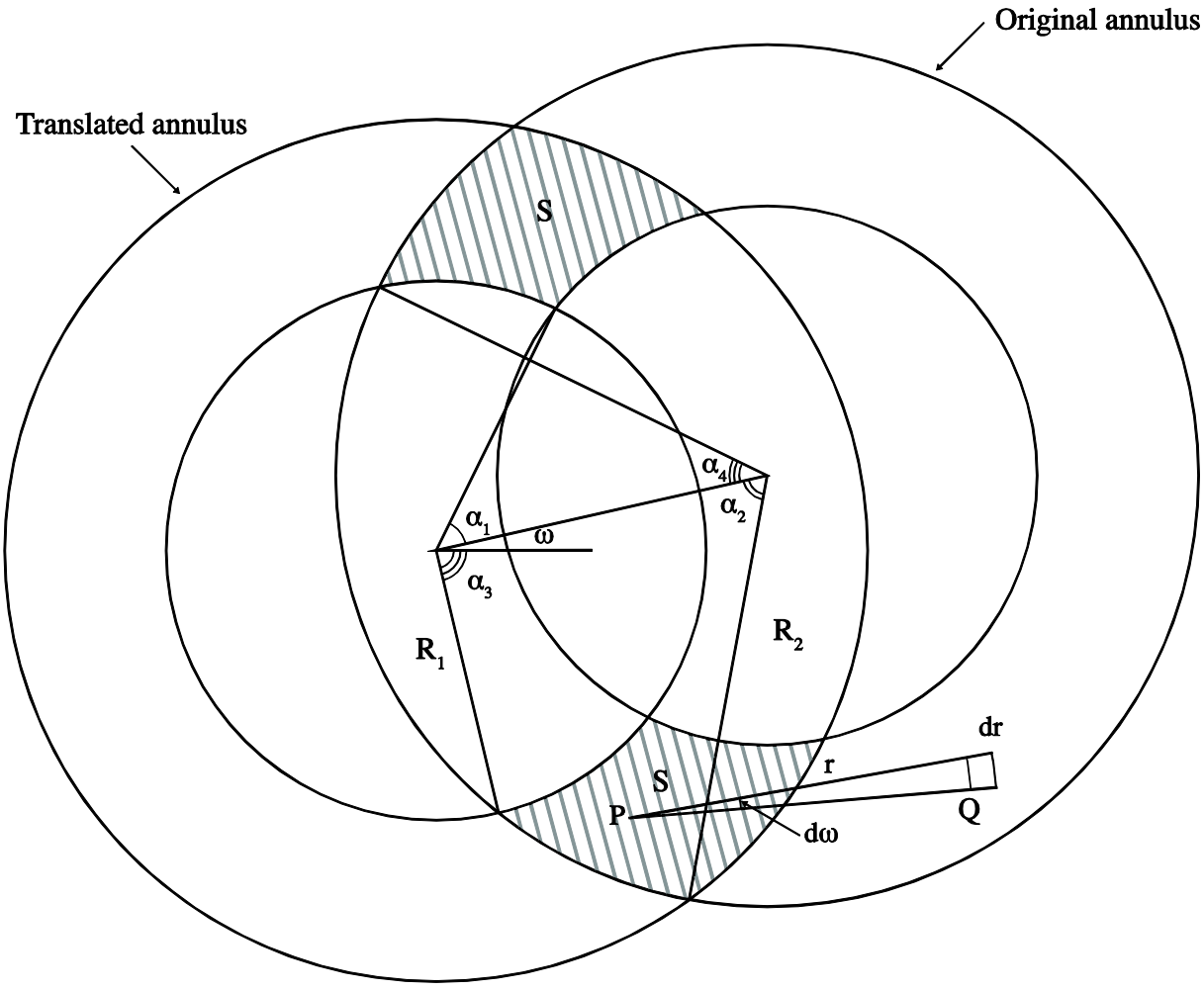


Figure 2. The annulus and its translated image, which enable us to define the areas S where P must lie, in order to define an acceptable (r, ω) pair.

The cosines of the angles shown in Figure 3 are: $\cos(\alpha_1) = \frac{r}{2R_1}$, $\cos(\alpha_2) = \frac{r}{2R_2}$, $\cos(\alpha_3) = \frac{r^2 + R_1^2 - R_2^2}{2rR_1}$, and $\cos(\alpha_4) = \frac{r^2 + R_2^2 - R_1^2}{2rR_2}$.

The joint probability of drawing two point P and Q that define an acceptable (r, ω) pair is

$$f(r)dr\,d\omega = \frac{S\,r\,dr\,d\omega}{\pi^2 R_1^2 R_2^2}. \tag{35}$$

We first estimate the values of S according to the value of r relative to $R_1, R_2, 2R_1$ if it is lower than R_2, $R_2 - R_1$ and $R_1 + R_2$. Then we integrate $f(r)\,dr\,d\omega$ with respect to ω. To give the result of this integration, it is convenient to define:

$$K(r) := \frac{4r}{\pi\left(R_2^2 - R_1^2\right)^2}. \tag{36}$$

If $R_1 \le \frac{R_2}{2}$, we go to the next case; otherwise, if $R_2 - R_1 \le r < 2R_1$, then

$$f(r) = K(r)\left(R_2^2(\alpha_2 - \cos\alpha_2 \sin\alpha_2) - R_1^2(\pi - \alpha_1 + \cos\alpha_1 \sin\alpha_1)\right). \tag{37}$$

If $R_2 - R_1 \le r < 2R_1$, then

$$f(r) = K(r)\left(R_2^2(\alpha_2 - \alpha_4 - \cos\alpha_2 \sin\alpha_1) - R_1^2\left(\frac{\alpha_1 - \alpha_2 - \cos\alpha_1 \sin\alpha_1 - \sin(\alpha_3 + \alpha_4)\sin(\alpha_3 - \alpha_1) - \sin(2\alpha_1)\sin(\alpha_1 - \alpha_4)}{\sin(\alpha_1 + \alpha_4)}\right)\right). \tag{38}$$

If $R_1 \le \frac{R_2}{2}$ and $R_2 - R_1 \le r < R_1 + R_2$, or if $R_1 > \frac{R_2}{2}$ and $2R_1 \le r < R_1 + R_2$, then

$$f(r) = K(r)\left(R_2^2\left(\alpha_2 - \alpha_4 - \cos\alpha_2 \sin\alpha_2 + \cos\alpha_4 \sin\alpha_4\right) - R_1^2\left(\alpha_3 - \cos\alpha_3 \sin\alpha_3\right)\right). \quad (39)$$

If $R_1 + R_2 \le r \le 2R_2$, then

$$f(r) = K(r)\left(\alpha_2 - \cos\alpha_2 \sin\alpha_2\right). \quad (40)$$

Figure 4 presents the curve obtained for $R_1 = 10$ and $R_2 = 30$, which we shall compare to the case of inhabited areas of Réunion Island (points indicated by a cross "x"). Why we traced a solid curve up to 52 km and a dotted line beyond 52 km will be explained in section 3.2. We shall not test for Taylor's law here, as the formulas are too complicated, but as shown in section 1, it holds true for all scalings cR_1,and cR_2, for $c > 0$.

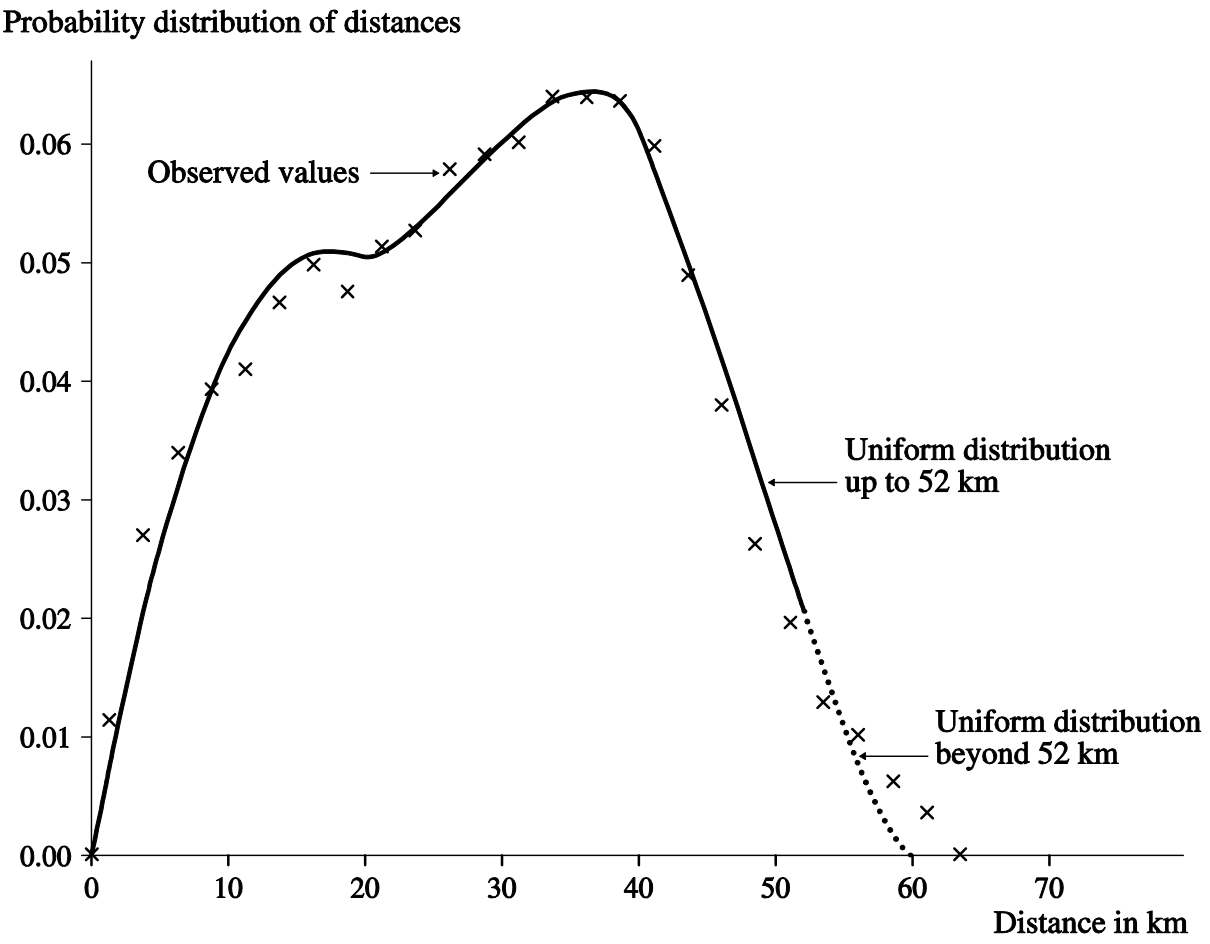


Figure 4. Probability distribution of distances obtained for an annulus of radii 10 and 30 with uniform distribution (solid line up to 52 km, dotted line beyond), with the values observed for inhabited areas of Réunion Island (points indicated by a cross "x").

2.5 Other shapes

Taylor's law applies to other shapes, but the formulas grow increasingly complicated from triangles (Borel, 1924) to rectangles (Mathai et al., 1999), convex polygons (Borel, 1924), and ellipsoids (Parry and Fischbach, 2000).

3. Populations distributed non-uniformly over a discretized territory

The distribution of distances between two random individuals in a continuous territory (such as those considered in section 2) could be analyzed assuming a non-uniform spatial distribution of population, by weighting the probability density function of the distance between the locations of the individuals by the population densities or counts at each terminus of the line segment between the two randomly chosen individuals (Tu and Fischbach, 2002). A computationally easy way to introduce population density into the calculations is to assume discrete populations with a finite total number of individuals located in a finite total number of areas. Most probability calculations should remain unchanged. Gridded population counts or estimates and gridded population maps are now available in many countries (at http://sedac.ciesin.columbia.edu/data/collection/gpw-v4). We can then verify theoretical results empirically.

3.1 Square populated by one individual in each cell

Consider a square of side R divided into n^2 square areas of side $\frac{R}{n}$, with a single individual at the center of each cell. The density of each such cell is $\delta = \frac{n^2}{R^2}$, with a total number of pairs of cells equal to

$$C = \frac{n^2\left(n^2-1\right)}{2}. \tag{41}$$

This number grows rapidly: for the $n^2 = 36{,}000$ French municipalities (approximately), the total number of pairs is nearly 6.5×10^8.

To count the total number of pairs at a distance $r = \frac{R\left(a^2+b^2\right)^{1/2}}{n}$, we shift the square of side R by one of the translation vectors $\overrightarrow{QP}$ $\left(a,\pm b\right)$ or $\left(b,\pm a\right)$, as we did with the annulus. The common surface of the square and each of its translated squares has an area equal to $\left(n-a\right)\left(n-b\right)$, which is also the total number of pairs for each translation. If $a=b$ or $a=0$ or $b=0$, then there are two possible translations, and four otherwise. The probability mass functions of the distance X between two points chosen at random in the square is equal to this count divided by the total number of cases given in Eq. 41.

We calculate the mean distance between two individuals selected at random on the territory

$$E\left(X\right) = \frac{2}{n^2\left(n^2-1\right)}\frac{R}{n}\left(2n\left(n-1\right)+2\sqrt{2}\left(n-1\right)^2+4n\left(n-2\right)+4\sqrt{5}\left(n-1\right)\left(n-2\right)+\cdots\right), \tag{42}$$

and its variance

$$\mathrm{Var}\left(X\right) = \frac{2}{n^2\left(n^2-1\right)}\frac{R^2}{n^2}\left(2n\left(n-1\right)+4\left(n-1\right)^2+8n\left(n-2\right)+20\left(n-1\right)\left(n-2\right)+\cdots\right). \quad \text{....(43)}$$

This distribution obeys Taylor's law with $b = 2$. However, the relationship between $E(X)$ and $\mathrm{Var}(X)$ now depends on the total number of areas. When the latter increases, the distribution tends toward a continuous distribution.

Figure 5 gives the example of a square of side $\frac{\sqrt{2}}{2}$ divided into 100 square areas of side $\frac{\sqrt{2}}{20}$, with distances grouped into 14 classes. The mean $E(X) = 0.5229$ and the variance $\mathrm{Var}(X) = 0.0628$ of the distance between a pair of randomly selected points from this distribution are very close to the values for a population distributed uniformly over the same square, for which $E(X) = 0.5214$ and $\mathrm{Var}(X) = 0.0615$. Figure 5 presents the fit.

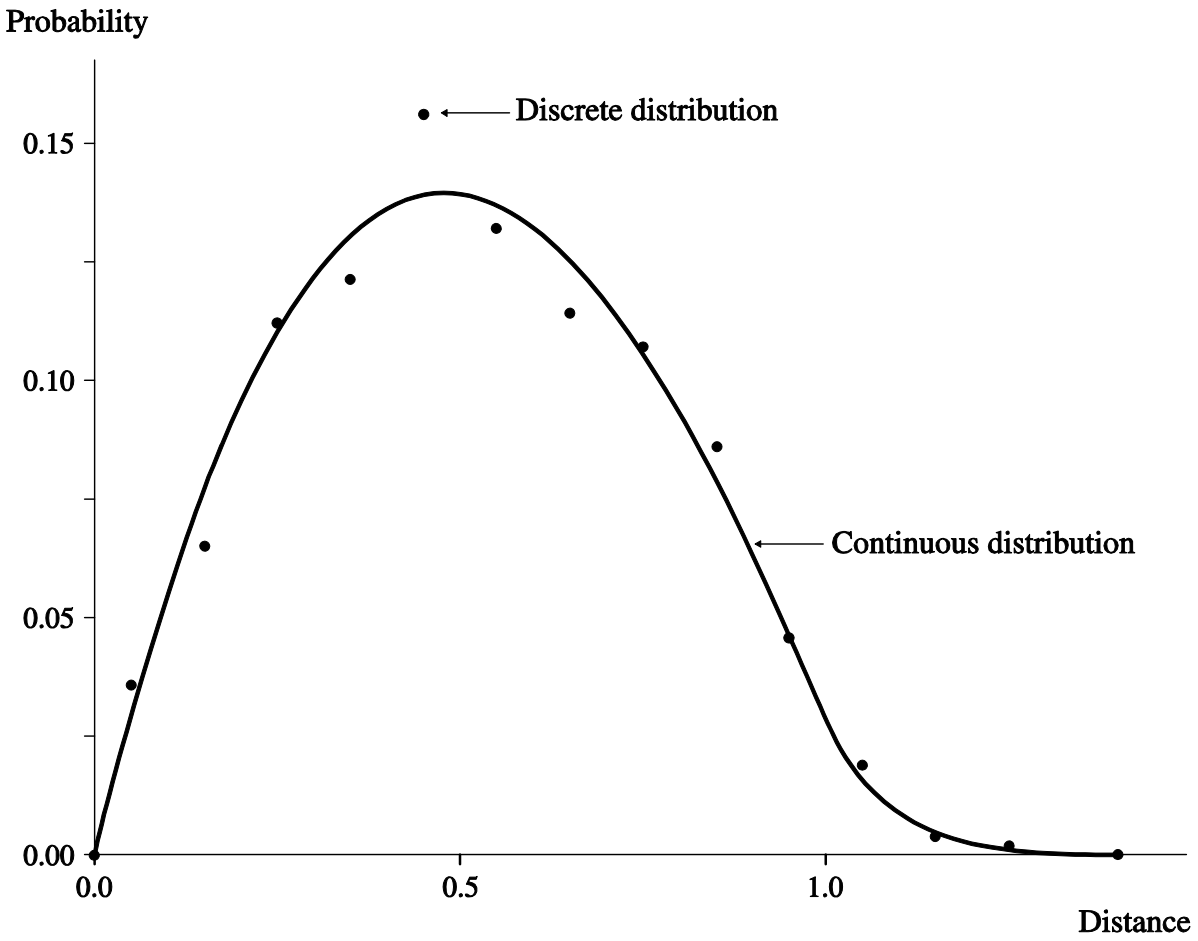


Figure 5. Distribution of distances grouped in 14 classes for a square of side $\sqrt{2}/2$ divided into squares of side $\frac{\sqrt{2}}{20}$ (black dots) compared with a continuous distribution (black line).

3.2 Réunion Island and France discretized by uniform gridded cells

Geographic grids of 1 square km introduced in several countries provide comparable data on human population distributions. The Lambert Azimuthal Equal Area projection (INSPIRE_Specification_GGS_v3.0.1.pdf) is the standard in Europe. The systems used in other areas are very similar.

Réunion

Réunion Island is roughly circular, with an uninhabited central region (Figure 2). Each cell of this annulus is associated with the indicator "inhabited or not". Using the data supplied by Insee, we calculate the distribution of distances between pairs of randomly chosen inhabited areas.

Figure 4 shows empirical along with theoretical probabilities of distances between random pairs of individuals resulting from the model of Eq. 37 to 40 for a uniform distribution on an annulus. The root mean squared error

$$\mathrm{RMSE} = \left(\frac{\sum_{i=1}^{n} (\hat{p}_i - p_i)^2}{n-k} \right)^{1/2}, \tag{44}$$

quantifies the goodness of fit, where $\hat{p}_i$ is the observed frequency for the distance d_i, p_i the theoretical frequency, n the total number of observed distances, and k the total number of estimated parameters. In Figure 1, the shape of the territory (a square or a circle) affects the tail of the distribution. We estimate the root mean squared error omitting this tail, which is

empirically determined when the theoretical and the observed curve begin to diverge, as the boundaries become more complicated. For Réunion Island, setting this tail at distances above 52 km leads to a root mean square error of 0.003025 for 19 degrees of freedom (21 classes of distance minus two estimated parameters), for an internal radius of 10 km and an external radius of 30 km. From 52 km to the maximum distance of 70 km (dotted line in Figure 3), the empirical distribution's tail is longer than for the annulus because the island is ellipsoidal rather than circular.

Metropolitan France

Metropolitan France comprises 375,279 inhabited areas of 1 square km, which makes nearly 70.4×10^9 pairs of areas. A 10% sample is enough and tractable to estimate the distribution of distances. Figure 6 compares the sampled frequency distribution of distances between two randomly selected inhabited square kilometers (indicated by crosses “x”) with the theoretical probability density function of distances between uniformly randomly chosen points from a square with side 728 km. We estimated the size of this square by a nonlinear least-square regression. The resulting root mean squared error was 0.00012 with 144 degrees of freedom, for distances less than 730 km.

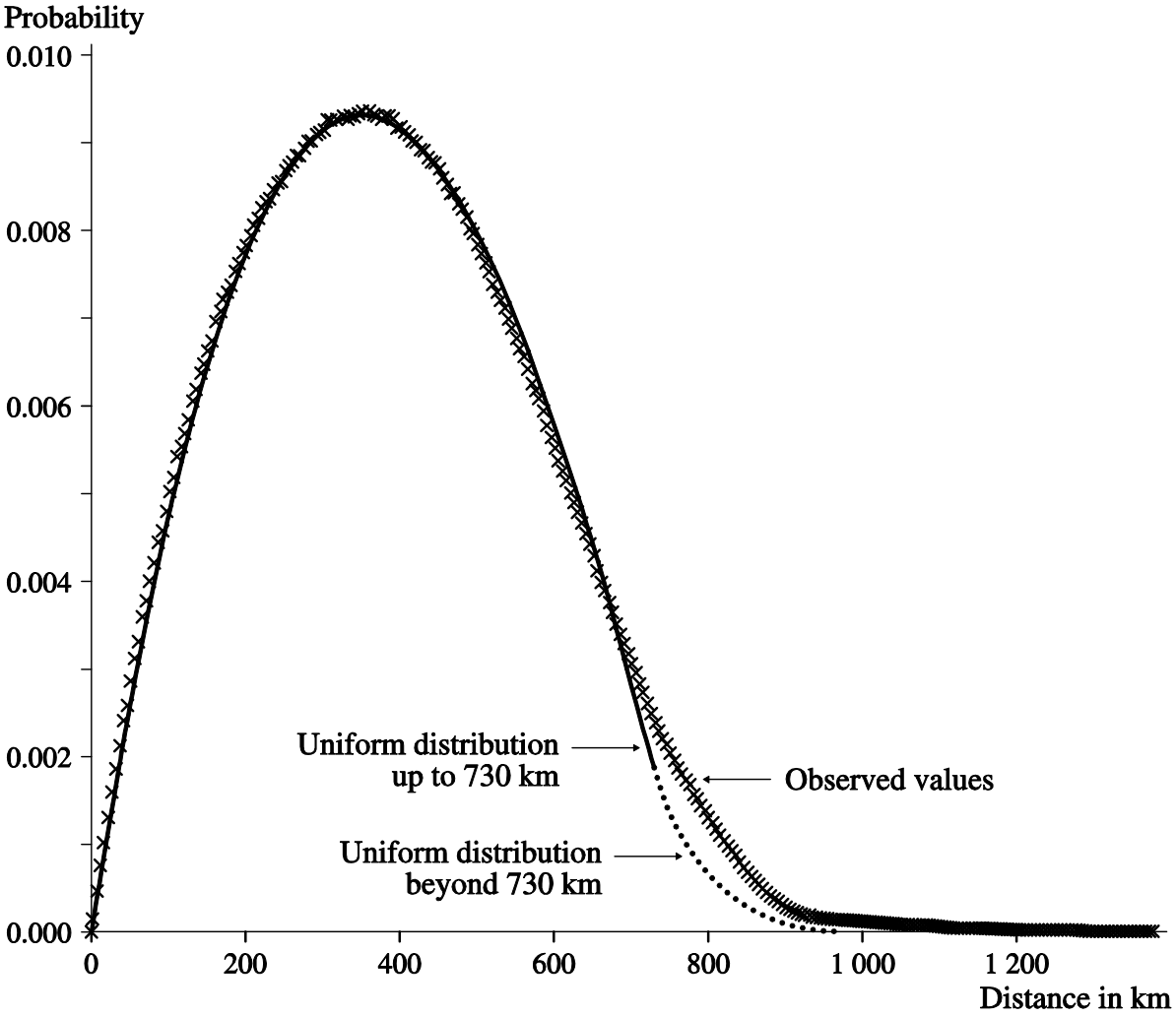


Figure 6. Sampled frequency distribution of distances (indicated by crosses "x") between 1 km × 1 km squares for France compared with the theoretical probability density function of distances between random pairs of points from a square with side 728 km (solid line up to 730 km, dotted line beyond).

Figure 6 shows that the empirical frequency exceeds the theoretical frequency for distances from 730 km to 1370 km (dotted line on Figure 6). The empirical frequency slightly exceeds the theoretical frequency for distances up to 200 km, and it is slightly less than the theoretical frequency for distances from 500 km to 650 km.

However, the surface of France is far from square, and only 375,279 of 551,500 areas of 1 square km are populated. The similarity of the sampled and the theoretical curves shows

that the presence of many uninhabited areas does not affect the center of the distribution of distances between inhabited areas.

3.3 Square with non-uniformly populated areas

We now consider non-uniformly populated areas. A square of side R is divided into n^2 equal square cells i with different population sizes, P_i, assuming that individuals are uniformly distributed in each cell. The country's total population is $P = \sum_{i=1}^{n^2} P_i$ and the population density of cell i is $\delta_i = P_i \frac{n^2}{R^2}$. The total number of pairs of individuals in the population is

$$C = \frac{P(P-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{n^2} P_i^2 + 2\sum_{j \neq i} \sum_{i=1}^{n^2} P_i P_j - \sum_{i=1}^{n^2} P_i \right). \quad (45)$$

A population of 65×10^6 individuals, as in France, counts 2.1×10^{15} pairs of individuals.

We determine the mean distance between individuals inhabiting the same cell. This case is similar to Eq. 33 for the continuous distribution, yielding a mean distance close to $\frac{R}{2n}$. The total number of pairs is given by the first and the third terms in Eq. 45:

$$\sum_{i=1}^{n^2} \frac{P_i(P_i-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{n^2} P_i^2 - \sum_{i=1}^{n^2} P_i \right). \quad (46)$$

The total number of pairs of individuals inhabiting different cells equals the sum of products $P_i P_j$ for each of the pairs of areas located at a given distance r.

3.4 Computing with population density for Réunion and metropolitan France

Réunion

Figure 7 is built from the gridded Insee data for Réunion Island on the population sizes of each area of 1 square km, which are the population densities.

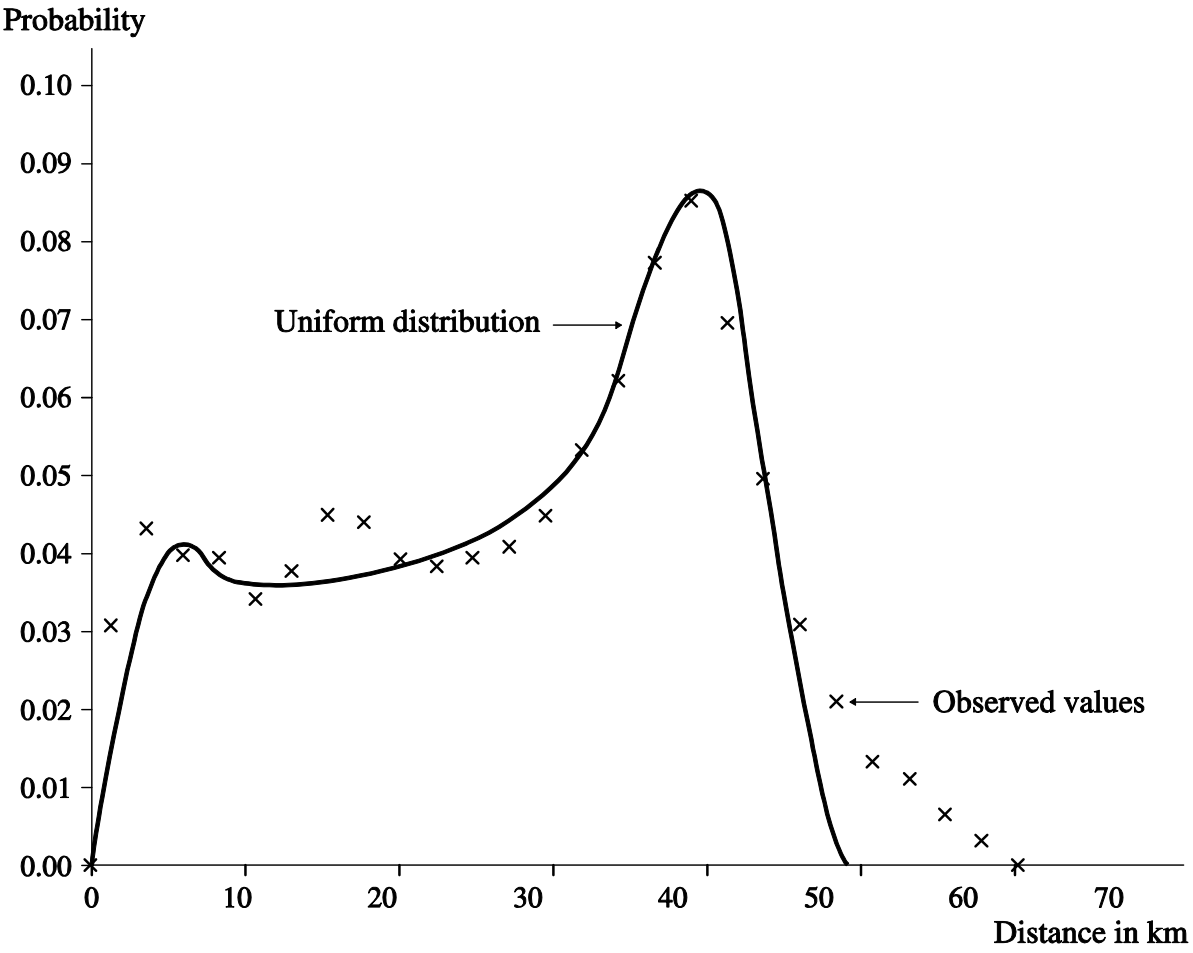


Figure 7. Frequency distribution of distances in km between pairs of individuals observed for Réunion Island (crosses) compared with the theoretical distribution obtained for an annulus (solid line) with uniform population density

As the population is concentrated in coastal cities, we used a smaller range of internal and external radius values, from 18 km to 26 km, than in our comparison between inhabited and uninhabited areas. Omitting distances greater than 52 km, the root mean squared error for 19 degrees of freedom is 0.007349, more than twice the previous value. The empirical distribution's tail is longer because the island is ellipsoidal, and has slightly higher observed

frequencies around 15 km, which is the distance between the island's two largest cities: Saint-Denis (population size 144,000) and Saint-Paul (population size 103,000).

Metropolitan France

The results for metropolitan France depend on the spatial scale. First we group distances into intervals of 100 km by aggregating the observations at the fine scale of 1 square km. We compare this distribution with the theoretical uniform distribution for a square (Figure 8).

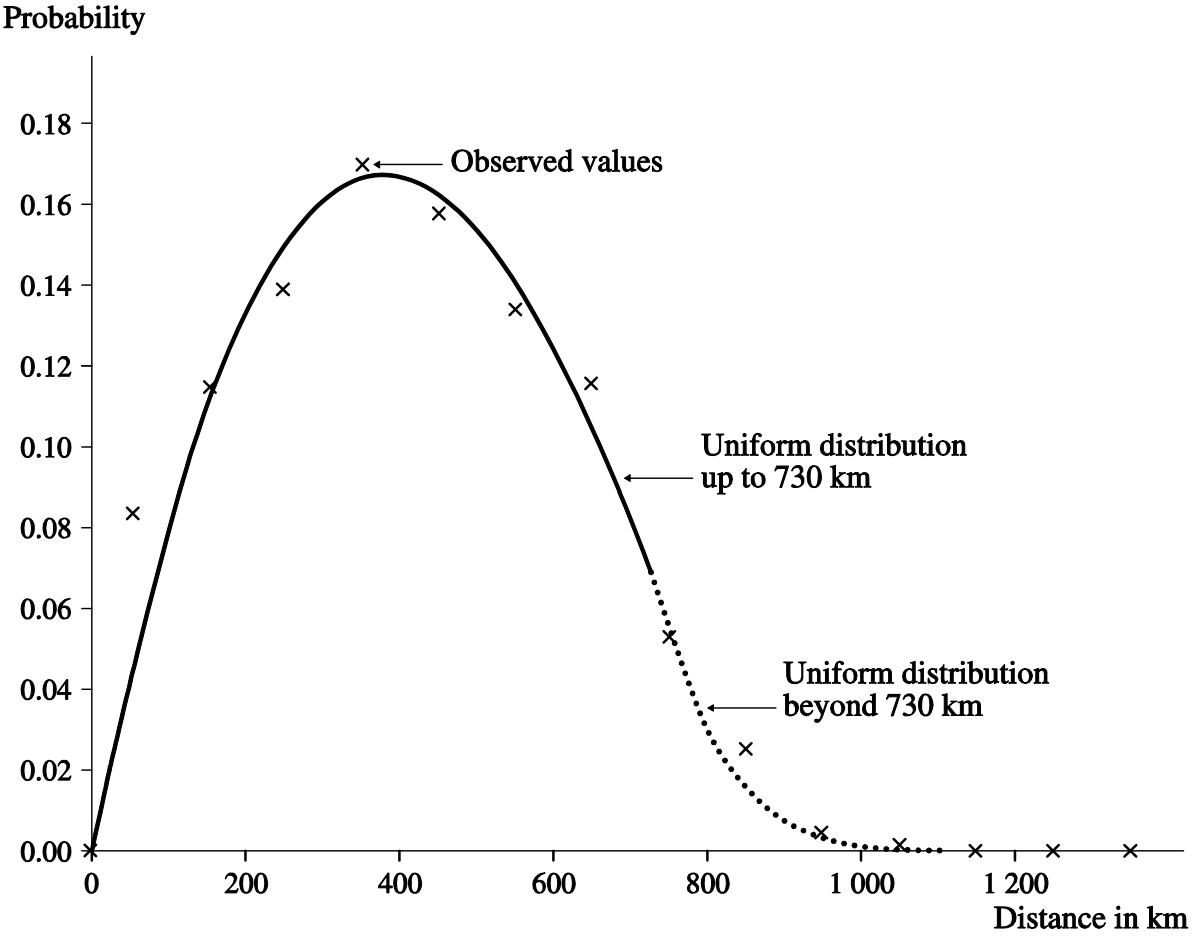


Figure 8. Frequency distribution of distances in km between pairs of randomly chosen individuals in France for a grouping into 100-km classes (indicated by crosses "x"), compared with the theoretical distribution (solid line up to 730 km, dotted line hereafter) with uniform population density on a square.

In this case, the root mean squared error is 0.01064 for 14 degrees of freedom, for a theoretical square with an estimated side of 791 km. The main discrepancy is for the first distance, where the observed frequency is more than twice the theoretical frequency. It is due to the population concentration in metropolitan areas.

If distances are grouped into 5-km classes and compared to a theoretical uniform distribution of population in a square with an estimated side of 798 km (Figure 9), the root mean squared error is 0.001212 for 144 degrees of freedom, ten times higher than the value calculated for inhabited areas. Deviations from the theoretical curve are noteworthy.

The local relative maxima at certain distances reflect the distribution of France's population living in large cities. They correspond to "as the crow flies" distances between pairs of regional metropolises: around 100 km between Paris and Rouen or between Paris and Orléans, 200 km between Paris and Lille, 300 km between Lyon and Marseille, 400 km between Paris and Lyon or between Paris and Strasbourg, 500 km between Paris and Bordeaux, 600 km between Paris and Toulouse, 700 km between Paris and Marseille. The peaks at short distance corresponds to the suburbs of these metropolises, in particular the Paris conurbation, as already observed on the 100-km scale (Figure 8). Réunion Island also displays a maximum for the distance between Saint-Denis and Saint-Paul. The smooth pattern predicted by geometric probability is overlaid by the discrete pattern of urban concentrations, with large cities at distances of multiples of 100 km in France.

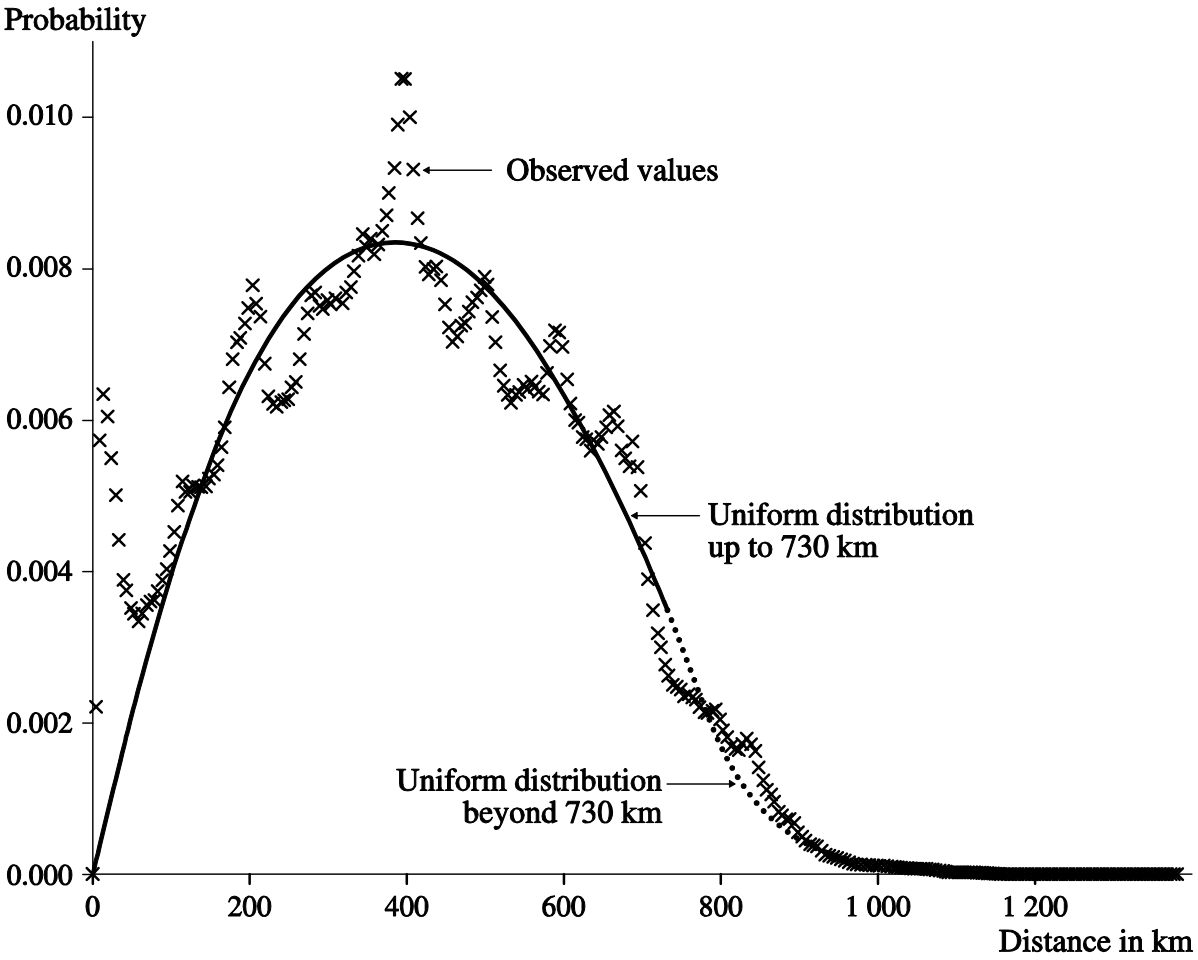


Figure 9. Frequency distribution of distances (km) between pairs of individuals in France for a grouping into 5-km classes (crosses), compared with the theoretical distribution (solid line up to 730 km, dotted line beyond) of a square with estimated side 798 km with uniform population density.

Conclusion

For elements distributed spatially, Taylor's law relates the mean and the variance of densities in sub-groups. We have shown theoretically that Taylor's law applies also to the distances between two randomly chosen points in various geometric shapes and under broad conditions.

For Réunion Island and metropolitan France and at some spatial scales, the empirical frequency distributions of inter-individual distances are predicted accurately by the theoretical frequency distributions of inter-point distances in models of geometric probability based on

the uniform distribution. When these models predict poorly, they provide baselines against which to highlight concentrations of population. For example, the observed distribution of distances of randomly drawn pairs of people in France is close to the theoretical distribution when distances are grouped into intervals of 100 km, but grouping distances into intervals of 5 km highlights peaks of observed frequency corresponding to the distances between regional metropolises.

ACKNOWLEDGMENTS

We thank Arnaud Bringé, head of Ined's Statistical Methods Department, for his very active involvement in the programming and calculations performed on Insee gridded data, and Cyril Courgeau for developing figures and map. We also thank Jonathan Mandelbaum for translating the section of this article initially drafted in French, and the editor Noel Bonneuil and an anonymous reviewer for many constructive suggestions and criticisms. J.E.C. thanks the U.S. National Science Foundation for grant DMS-1225529 and Priscilla K. Rogerson for assistance.

References

Barton, D.E., David, F. N., and Fix, E. (1963). Random points in a circle and the analysis of chromosome patterns. *Biometrika*, *50*(1/2): 23-29.

Bell, M., Elin, C.-E., Ueffing, P., Stillwell, J., Kupiszewski, M., and Kupiszevka D. (2015). Internal migration and development: comparing migration intensities around the world. *Population and Development Review*, *41*(1): 33-58.

Bénard, F. (2012), *Formes urbaines et transport en milieu insulaire : l'exemple de la Réunion*. Saint Denis: Université de la Réunion, <NNT : 2012LARE0020>

Borel, E. (1924). *Principes et formules classiques du calcul des probabilités.* Paris: Gauthier-Villars.

Boutin, M. and Kemper; G. (2004). On reconstructing *n*-point configurations from the distribution of distances or areas. *Advances in Applied Mathematics*, *32*(4): 709-735.

Buffon (de), G.-L. (1733). Solutions de problèmes qui regardaient le jeu du franc carreau, Résumé by Fontenelle B., in *Histoire de l'Académie Royale des Sciences de Paris*, Boudot J. (ed). Paris: Imprimerie Royale, 43-45.

Cohen, J. E., Xu, M., and Brunborg H. (2013). Taylor's law applies to spatial variation in a human population. *Genus*, *29*(1): 25-60.

Courgeau, D. (1970). *Les champs migratoires en France*. Paris: Presses Universitaires de France.

Courgeau, D. (1973). Migrations et découpages du territoire. *Population*, *28*(3): 511-537.

Courgeau, D. and Baccaïni, B. (1989). Migrations et distances. *Population*, *42*(1): 57-82.

Crofton, M. W. (1885). Probability, in *Encyclopaedia Britannica* (9th Edition, vol. *19*), T.S. Baynes (ed). Edinburgh: A & C Black, 768-788.

Garwood, F. (1947). The variance of the overlap of geometrical figures with reference to a bombing problem. *Biometrika*, *34*(1/2): 1-17.

Giometto, A., Formentin, M., Rinaldo, A., Cohen, J. E., and Maritan, A. (2015). Sample and population exponents of generalized Taylor's law. *Proceedings of the National Academy of Sciences 112*(25): 7755-7760.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. New York: Wiley.

Kuiper, J. H. and Paelinck J. H. (1982). Frequency distribution of distances and related concepts. *Geographical Analysis*, *14*(3): 253-259.

Luu Mau Thanh (1962). Distribution théorique des distances entre deux points répartis uniformément sur le territoire, in *Les déplacements humains*, Entretiens de Monaco en sciences humaines, Sutter J. (ed). Paris: Hachette, 173-184.

Mathai, A. M., Moschopoulos, P., and Pederzoli, G. (1999). Random points associated with rectangles. *Rendiconti del Circolo Matematico Di Palermo*, *48*(1): 162-190.

Moltchanov, D. (2012). Distance distributions in random networks. *Ad Hoc Networks*, *10*(6): 1146-1166.

Parry, M. and Fischbach, E. (2000). Probability distribution of distance in a uniform ellipsoid: theory and applications to physics. *Journal of Mathematical Physics*, *41*(4): 2417-2433.

Ramsayer, J., Fellous, S., Cohen, J. E., and Hochberg, M. E. (2012). Taylor's law holds in experimental bacterial populations but competition does not influence the slope. *Biology Letters, 8*(2): 316-319.

Rogerson, P. (1990). Buffon's needle and the estimation of migration distances. *Mathematical Population Studies*, *2*(3): 229-238.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge, New York: Cambridge University Press.

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, *189*: 732-735.

Taylor, L. R., Woiwod I. P., and Perry J. N. (1978). Density dependence of spatial behaviour and the rarity of randomness. *Journal of Animal Ecology*, *47*(2): 383-406.

Tu, S.-J., and Fischbach, E. (2002). Random distance distribution for spherical objects: general theory and applications to physics. *Journal of Physics A: Mathematical and General*, *35*(31): 6557-6570.

Uttley, P. and M^{c}Hardy I. M. (2001). The flux-dependent amplitude of Broadband noise in X-ray binaries and active galaxies. *Monthly Notes of the Royal Astronomical Society*, *323*: L26-L30.