Simulating Dependent Discrete Data

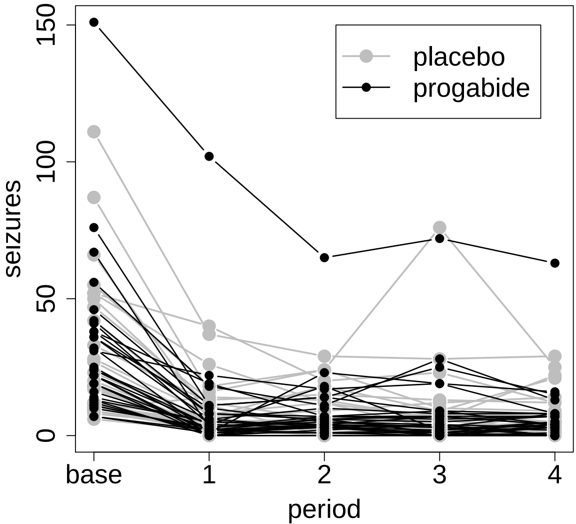
Lisa Madsen
Dave Birkes

Portland State University
January 17, 2014

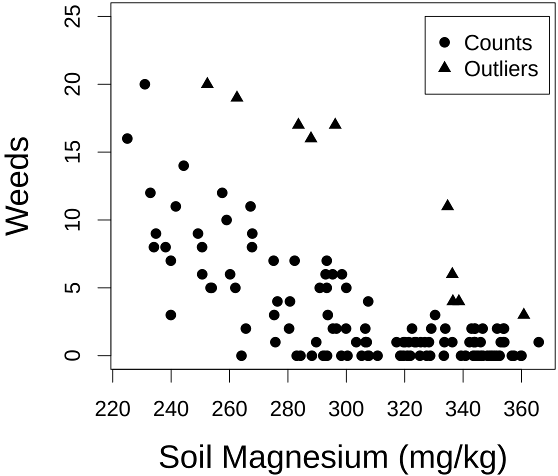
Outline

Outline

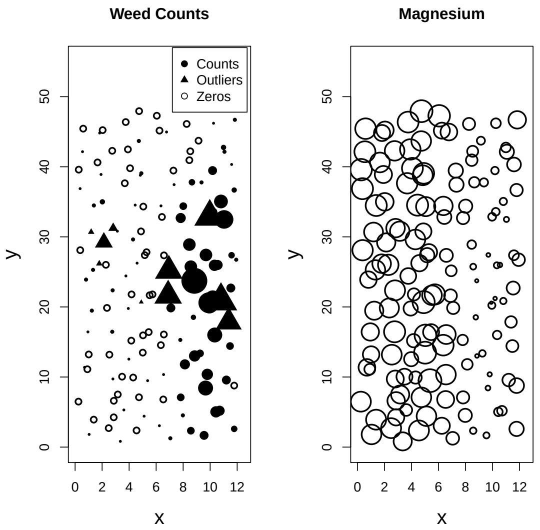
Seizure Counts Over Time (Diggle et al., 2002)



Weed Counts vs. Soil Magnesium (Heijting et al, 2007)



Maps of Weed Counts and Magnesium



Outline

Why Simulate Data?

- Assess the performance of analytical procedures

Why Simulate Data?

- Assess the performance of analytical procedures
- Compare two or more statistical methods

Why Simulate Data?

- Assess the performance of analytical procedures
- Compare two or more statistical methods
- Parametric bootstrap, e.g. for goodness of fit tests

Why Simulate Data?

- Assess the performance of analytical procedures
- Compare two or more statistical methods
- Parametric bootstrap, e.g. for goodness of fit tests
- Power analysis or sample size determination

Why Simulate Data?

- Assess the performance of analytical procedures
- Compare two or more statistical methods
- Parametric bootstrap, e.g. for goodness of fit tests
- Power analysis or sample size determination
- Find a good sampling design

Outline

Pearson Correlation

The usual measure of dependence between X and Y is the Pearson product-moment correlation coefficient:

$$\rho(X, Y) = \frac{E\{[X - E(X)][Y - E(Y)]\}}{[\mathrm{var}(X)\,\mathrm{var}(Y)]^{1/2}} = \frac{E(XY) - E(X)E(Y)}{[\mathrm{var}(X)\,\mathrm{var}(Y)]^{1/2}}.$$

Pearson Correlation

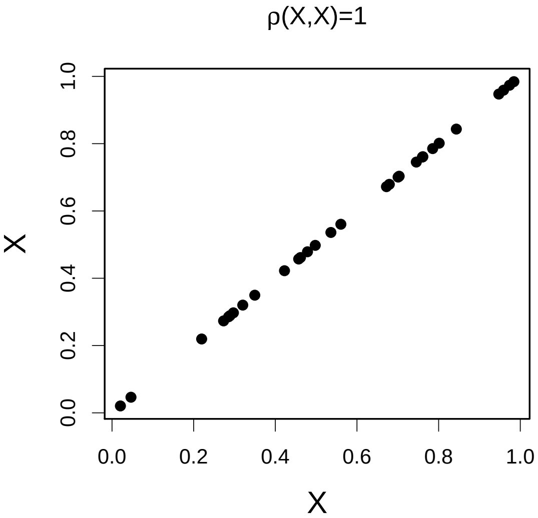
The usual measure of dependence between X and Y is the Pearson product-moment correlation coefficient:

$$\rho(X,Y) = \frac{E\{[X-E(X)][Y-E(Y)]\}}{[\mathrm{var}(X)\,\mathrm{var}(Y)]^{1/2}} = \frac{E(XY)-E(X)E(Y)}{[\mathrm{var}(X)\,\mathrm{var}(Y)]^{1/2}}.$$

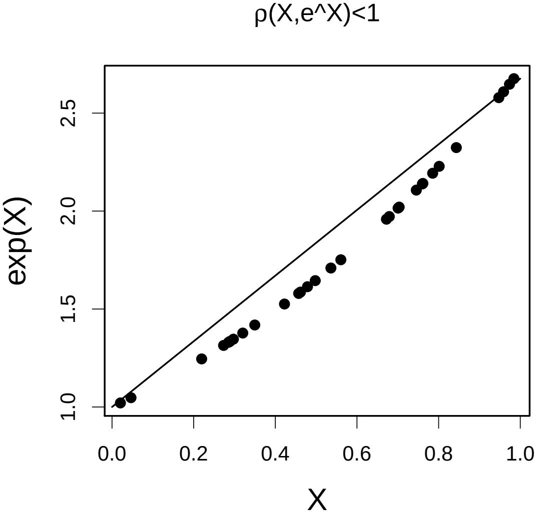
Estimate $\rho(X,Y)$ from sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ as

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{n}[(X_i-\overline{X})(Y_i-\overline{Y})]}{[\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2]^{1/2}},$$

Pearson Correlation Measures Linear Dependence



Pearson Correlation Measures Linear Dependence



Pearson Correlation Measures Linear Dependence

For bivariate normal X and Y, $\rho(X, Y)$ completely characterizes dependence.

Pearson Correlation Measures Linear Dependence

For bivariate normal X and Y, $\rho(X, Y)$ completely characterizes dependence.

For non-normal X and Y, other measures of dependence may be more appropriate.

Outline

Spearman Correlation

The Spearman correlation coefficient is

$$\rho_S(X,Y) = 3\{P[(X-X_0)(Y-Y_0) > 0] - P[(X-X_0)(Y-Y_0) < 0]\}$$

where

$$\begin{aligned} X_0 &\stackrel{d}{=} X \\ Y_0 &\stackrel{d}{=} Y \end{aligned}$$

with X_0 and Y_0 independent of one another and of (X, Y).

Estimating Spearman Correlation

Given bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, calculate ranks $r(X_i)$ and $r(Y_i)$. Then

$$\hat{\rho}_S(X, Y) = \frac{\sum_{i=1}^{n}\{[r(X_i) - (n+1)/2][r(Y_i) - (n+1)/2]\}}{n(n^2-1)/12},$$

the sample Pearson correlation coefficient of the ranked data.

Example of Ranked Bivariate Sample

$$(X_1, Y_1), \ldots, (X_n, Y_n) = (1, 5), (3, 3), (0, 2), (5, 4)$$

Example of Ranked Bivariate Sample

$$(X_1, Y_1), \ldots, (X_n, Y_n) = (1, 5), (3, 3), (0, 2), (5, 4)$$

Ordered X's: $0, 1, 3, 5$

Example of Ranked Bivariate Sample

$$(X_1, Y_1), \ldots, (X_n, Y_n) = (1,5), (3,3), (0,2), (5,4)$$

Ordered X's: $0, 1, 3, 5$

Ordered Y's: $2, 3, 4, 5$

Example of Ranked Bivariate Sample

$$(X_1, Y_1), \ldots, (X_n, Y_n) = (1,5), (3,3), (0,2), (5,4)$$

Ordered X's: $0, 1, 3, 5$

Ordered Y's: $2, 3, 4, 5$

Rank is position in ordered list:

$$[r(X_1), r(Y_1)], \ldots, [r(X_n), r(Y_n)] = (2,4), (3,2), (1,1), (4,3).$$

Spearman Correlation Measures Monotone Dependence

$$\rho_S(X, e^X) = \rho_S(X, X) = 1 \ldots$$

Spearman Correlation Measures Monotone Dependence

$\rho_S(X, e^X) = \rho_S(X, X) = 1 \ldots$ provided X is continuous.

Correcting for Ties

When X is discrete, it is possible to have X and Y so that $X = Y$ almost surely but $\rho_S(X, Y) < 1$.

Correcting for Ties

When X is discrete, it is possible to have X and Y so that $X = Y$ almost surely but $\rho_S(X, Y) < 1$.

Rescale ρ_S so that it ranges between -1 and 1:

$$\rho_{RS}(X, Y) = \frac{\rho_S(X, Y)}{\{[1 - \sum_x p(x)^3][1 - \sum_y q(y)^3]\}^{1/2}},$$

where $p(x) = P(X = x)$ and $q(y) = P(Y = y)$ (Nešlehová, 2007).

Ties in Sample Ranks

Two common methods for handling ties in sample $X_1, \ldots, X_n$:

- Random ranks: When u tied values would occupy ranks $p_1, \ldots, p_u$ if they were distinct, randomly assign these u ranks to the tied values.

Ties in Sample Ranks

Two common methods for handling ties in sample $X_1, \ldots, X_n$:

- Random ranks: When u tied values would occupy ranks $p_1, \ldots, p_u$ if they were distinct, randomly assign these u ranks to the tied values.

$$0, 8, 4, 4, 4 \;\rightarrow 1, 5, c_1, c_2, c_3$$

where c_1, c_2, c_3 is a random permutation of $2, 3, 4$.

Ties in Sample Ranks

Two common methods for handling ties in sample $X_1, \ldots, X_n$:

- Random ranks: When u tied values would occupy ranks $p_1, \ldots, p_u$ if they were distinct, randomly assign these u ranks to the tied values.

$$0, 8, 4, 4, 4 \ \rightarrow 1, 5, c_1, c_2, c_3$$

  where c_1, c_2, c_3 is a random permutation of $2, 3, 4$.

- Midranks: Assign each tied value the average rank, $\frac{1}{u}\sum_{k=1}^{u} p_k$.

Ties in Sample Ranks

Two common methods for handling ties in sample $X_1, \ldots, X_n$:

- Random ranks: When u tied values would occupy ranks $p_1, \ldots, p_u$ if they were distinct, randomly assign these u ranks to the tied values.

$$0, 8, 4, 4, 4 \ \rightarrow 1, 5, c_1, c_2, c_3$$

  where c_1, c_2, c_3 is a random permutation of $2, 3, 4$.

- Midranks: Assign each tied value the average rank, $\frac{1}{u} \sum_{k=1}^{u} p_k$.

$$0, 8, 4, 4, 4 \ \rightarrow 1, 5, 3, 3, 3$$

Rescaled Spearman Correlation and Midranks

For sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, let the distribution of (X, Y) be the empirical distribution function of the sample. Then $\rho_{RS}(X, Y)$ coincides with the sample Pearson correlation coefficient of the midranks (Nešlehová, 2007).

Outline

Fréchet-Hoeffding Bounds

For X and Y with joint CDF $H(x, y)$ and marginal CDFs $F(x)$ and $G(y)$, the Fréchet-Hoeffding bounds are

$$\max[F(x) + G(y) - 1, 0] \leq H(x, y) \leq \min[F(x), G(y)]$$

Fréchet-Hoeffding Bounds

For X and Y with joint CDF $H(x, y)$ and marginal CDFs $F(x)$ and $G(y)$, the Fréchet-Hoeffding bounds are

$$\underbrace{\max[F(x) + G(y) - 1, 0]}_{W[F(x), G(y)]} \leq H(x, y) \leq \underbrace{\min[F(x), G(y)]}_{M[F(x), G(y)]}.$$

Fréchet-Hoeffding Bounds

For X and Y with joint CDF $H(x, y)$ and marginal CDFs $F(x)$ and $G(y)$, the Fréchet-Hoeffding bounds are

$$\underbrace{\max[F(x) + G(y) - 1, 0]}_{W[F(x),G(y)]} \leq H(x, y) \leq \underbrace{\min[F(x), G(y)]}_{M[F(x),G(y)]}.$$

These bounds induce margin-dependent bounds on $\rho(X, Y)$ and $\rho_S(X, Y)$:

Fréchet-Hoeffding Bounds

For X and Y with joint CDF $H(x, y)$ and marginal CDFs $F(x)$ and $G(y)$, the Fréchet-Hoeffding bounds are

$$\underbrace{\max[F(x) + G(y) - 1, 0]}_{W[F(x), G(y)]} \leq H(x, y) \leq \underbrace{\min[F(x), G(y)]}_{M[F(x), G(y)]}.$$

These bounds induce margin-dependent bounds on $\rho(X, Y)$ and $\rho_S(X, Y)$:

$$\rho\{W[F(x), G(y)]\} \quad \leq \quad \rho(X, Y) \quad \leq \quad \rho\{M[F(x), G(y)]\}$$

Fréchet-Hoeffding Bounds

For X and Y with joint CDF $H(x, y)$ and marginal CDFs $F(x)$ and $G(y)$, the Fréchet-Hoeffding bounds are

$$\underbrace{\max[F(x) + G(y) - 1, 0]}_{W[F(x), G(y)]} \leq H(x, y) \leq \underbrace{\min[F(x), G(y)]}_{M[F(x), G(y)]}.$$

These bounds induce margin-dependent bounds on $\rho(X, Y)$ and $\rho_S(X, Y)$:

$$\begin{aligned} \rho\{W[F(x), G(y)]\} &\leq \rho(X, Y) \leq \rho\{M[F(x), G(y)]\} \\ \rho_S\{W[F(x), G(y)]\} &\leq \rho_S(X, Y) \leq \rho_S\{M[F(x), G(y)]\} \end{aligned}$$

Outline

Simulation Algorithm

Suppose we want to simulate dependent $\boldsymbol{Y} = [Y_1, \ldots, Y_N]'$ where Y_i has marginal CDF F_i.

1. Simulate a multivariate standard normal vector $\boldsymbol{Z}$ with variance-covariance matrix $\boldsymbol{\Sigma_Z}$. Note: $\{\boldsymbol{\Sigma_Z}\}_{ij} = \rho(Z_i, Z_j)$.

Simulation Algorithm

Suppose we want to simulate dependent $\boldsymbol{Y} = [Y_1, \ldots, Y_N]'$ where Y_i has marginal CDF F_i.

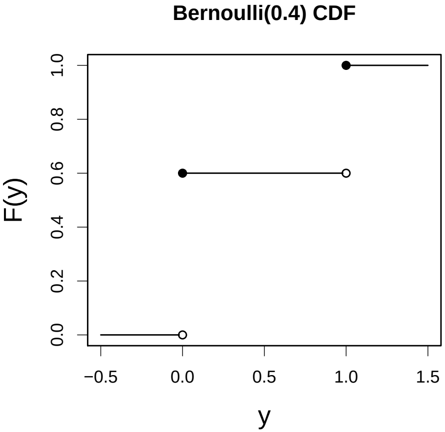
1. Simulate a multivariate standard normal vector $\boldsymbol{Z}$ with variance-covariance matrix $\boldsymbol{\Sigma_Z}$. Note: $\{\boldsymbol{\Sigma_Z}\}_{ij} = \rho(Z_i, Z_j)$.

2. Transform each element of $\boldsymbol{Z}$ to obtain desired marginals:

$$Y_i = F_i^{-1}\{\Phi(Z_i)\},$$

where $\Phi(\cdot)$ denotes the standard normal CDF.

Inverse CDF for Discrete Distributions



$$F_i^{-1}(u) = \inf\{y : F_i(y) \geq u\}$$

$\mathrm{corr}(Z_i, Z_j) \neq 0$ Induces Dependence Between Y_i, Y_j

Since $Y_i = F_i^{-1}\{\Phi(Z_i)\}$, both $\rho(Y_i, Y_j)$ and $\rho_S(Y_i, Y_j)$ can be written as functions of F_i, F_j, and $\rho(Z_i, Z_j)$.

$\text{corr}(Z_i, Z_j) \neq 0$ Induces Dependence Between Y_i, Y_j

Since $Y_i = F_i^{-1}\{\Phi(Z_i)\}$, both $\rho(Y_i, Y_j)$ and $\rho_S(Y_i, Y_j)$ can be written as functions of F_i, F_j, and $\rho(Z_i, Z_j)$.

Given target marginals F_i, F_j, and either $\rho(Y_i, Y_j)$ or $\rho_S(Y_i, Y_j)$, can numerically solve an equation to find $\rho(Z_i, Z_j)$.

Outline

Method Achieves Any ρ Within Fréchet-Hoeffding Bounds

Theorem 1

Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the described method. Assume Y_1 and Y_2 have finite variance. Let $\rho^*(\delta)$ denote $\rho(Y_1, Y_2)$ as a function of $\delta \equiv \rho(Z_1, Z_2)$. Then $\{\rho^*(\delta) : \delta \in [-1, 1]\} = [\rho(W), \rho(M)]$.

Proof.

ρ^* is a continuous function of δ and $\rho^*(-1) = \rho(W)$ and $\rho^*(-1) = \rho(M)$. $\square$

Method Achieves Any ρ Within Fréchet-Hoeffding Bounds

Theorem 1

Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the described method. Assume Y_1 and Y_2 have finite variance. Let $\rho^*(\delta)$ denote $\rho(Y_1, Y_2)$ as a function of $\delta \equiv \rho(Z_1, Z_2)$. Then $\{\rho^*(\delta) : \delta \in [-1, 1]\} = [\rho(W), \rho(M)]$.

Proof.

ρ^* is a continuous function of δ and $\rho^*(-1) = \rho(W)$ and $\rho^*(-1) = \rho(M)$. $\square$

Method Achieves Any ρ_S Within Fréchet-Hoeffding Bounds

Theorem 2

Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the described method. Assume F_1 and F_2 satisfy $\lim_{x \uparrow x_0} F_i(x) = F_i(x_0 - \epsilon_i)$ for all x_0 in the support of F_i, for some ϵ_i depending on F_i but not on x_0. Let $\rho^*_S(\delta)$ denote $\rho_S(Y_1, Y_2)$ as a function of δ. Then $\{\rho^*_S(\delta) : \delta \in [-1, 1]\} = [\rho_S(W), \rho_S(M)]$.

Proof.

ρ^*_S is a continuous function of δ and $\rho^*_S(-1) = \rho_S(W)$ and $\rho^*_S(-1) = \rho_S(M)$. □

Method Achieves Any ρ_S Within Fréchet-Hoeffding Bounds

Theorem 2

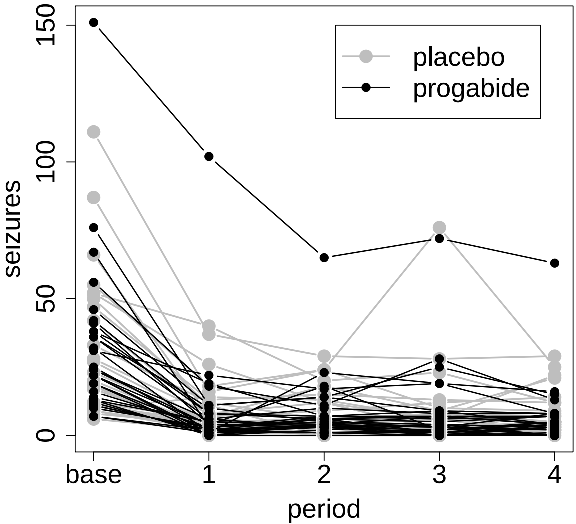
Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ denote a pair of random variables simulated according to the described method. Assume F_1 and F_2 satisfy $\lim_{x \uparrow x_0} F_i(x) = F_i(x_0 - \epsilon_i)$ for all x_0 in the support of F_i, for some ϵ_i depending on F_i but not on x_0. Let $\rho_S^*(\delta)$ denote $\rho_S(Y_1, Y_2)$ as a function of δ. Then $\{\rho_S^*(\delta) : \delta \in [-1, 1]\} = [\rho_S(W), \rho_S(M)]$.

Proof.

ρ_S^* is a continuous function of δ and $\rho_S^*(-1) = \rho_S(W)$ and $\rho_S^*(-1) = \rho_S(M)$. □

Outline

Seizure Data



Marginal Model (Diggle, et al. 2002)

Y_{ij} denotes jth observation on ith subject, $i = 1, \dots, 58$, $j = 0, \dots, 4$.

$$\mu_{ij} = E(Y_{ij}) = \exp[\log(t_j) + \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2i} + \beta_3 x_{1j} x_{2i}]$$

where

$$x_{1j} = \begin{cases} 0 & \text{if } j = 0 \text{ (baseline)} \\ 1 & \text{if } j = 1, 2, 3, \text{ or } 4 \end{cases}$$

$$x_{2i} = \begin{cases} 0 & \text{subject } i \text{ in placebo group} \\ 1 & \text{subject } i \text{ in progabide group} \end{cases}$$

$$t_j = \begin{cases} 8 & \text{if } j = 0 \\ 2 & \text{if } j = 1, 2, 3, \text{ or } 4 \end{cases}$$

$\sigma^2_{ij} = \text{var}(Y_{ij}) = \phi \cdot \mu_{ij}$

Marginal Model Parameter Estimates

Diggle et al. (2002) use Generalized Estimating Equation methodology to estimate model parameters. Plug estimates into model:

$$\begin{aligned} \widehat{\mu}_{ij} &= \exp[\log(t_j) + 1.35 + 0.11x_{1j} - 0.11x_{2i} - 0.3x_{1j}x_{2i}] \\ \widehat{\phi} &= 10.4 \end{aligned}$$

Marginal Model Parameter Estimates

Diggle et al. (2002) use Generalized Estimating Equation methodology to estimate model parameters. Plug estimates into model:

$$\begin{aligned} \widehat{\mu}_{ij} &= \exp[\log(t_j) + 1.35 + 0.11x_{1j} - 0.11x_{2i} - 0.3x_{1j}x_{2i}] \\ \widehat{\phi} &= 10.4 \end{aligned}$$

$\widehat{\phi} > 1 \Rightarrow$ overdispersed counts, e.g. negative binomial.

Marginal Model Parameter Estimates

Diggle et al. (2002) use Generalized Estimating Equation methodology to estimate model parameters. Plug estimates into model:

$$
\begin{aligned}
\widehat{\mu}_{ij} &= \exp[\log(t_j) + 1.35 + 0.11x_{1j} - 0.11x_{2i} - 0.3x_{1j}x_{2i}] \\
\widehat{\phi} &= 10.4
\end{aligned}
$$

$\widehat{\phi} > 1 \Rightarrow$ overdispersed counts, e.g. negative binomial.

Let $\widehat{F}_{ij}$ be the negative binomial CDF with mean $\widehat{\mu}_{ij}$ and variance $\widehat{\phi} \cdot \widehat{\mu}_{ij}$. These will be our target marginals.

Pearson Correlation

Diggle, et al. (2002) model dependence as

$$\rho(Y_{ij}, Y_{i'j'}) = \begin{cases} 0 & \text{if } i \neq i' \text{ (different subjects)} \\ \alpha & \text{if } i = i' \text{ and } j \neq j' \\ 1 & \text{if } i = i' \text{ and } j = j' \end{cases}$$

and calculate $\widehat{\alpha} = 0.6$.

Calculating $\boldsymbol{\Sigma}_Z$

For each pair $(Y_{ij}, Y_{ij'})$, $j \neq j'$, numerically solve for $\delta = \rho(Z_{ij}, Z_{ij'})$:

$$\hat{\rho}(Y_{ij}, Y_{ij'}) = \frac{1}{\widehat{\sigma}_{ij}\widehat{\sigma}_{ij'}}\left\{ \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\Big(1 - \widehat{F}_{ij}(r) - \widehat{F}_{ij'}(s) + \Phi_{\delta}\{\Phi^{-1}[\widehat{F}_{ij}(r)], \Phi^{-1}[\widehat{F}_{ij'}(s)]\}\Big) - \widehat{\mu}_{ij}\widehat{\mu}_{ij'} \right\}$$

where Φ_δ denotes the bivariate standard normal CDF with correlation δ.

Simulating Seizure Data

Apply algorithm:

1. Simulate multivariate standard normal vector $\boldsymbol{Z}$ with variance-covariance matrix $\boldsymbol{\Sigma_Z}$ where the elements of $\boldsymbol{\Sigma_Z}$ are either 0, 1, or solutions for δ to the equation corresponding to the pair $(Y_{ij}, Y_{ij'})$.

Simulating Seizure Data

Apply algorithm:

1. Simulate multivariate standard normal vector $\boldsymbol{Z}$ with variance-covariance matrix $\boldsymbol{\Sigma_Z}$ where the elements of $\boldsymbol{\Sigma_Z}$ are either 0, 1, or solutions for δ to the equation corresponding to the pair $(Y_{ij}, Y_{ij'})$.

2. Transform each element of $\boldsymbol{Z}$ to obtain desired marginals:

$$Y_{ij} = \widehat{F}_{ij}^{-1}\{\Phi(Z_{ij})\}$$

Simulating Seizure Data

Apply algorithm:

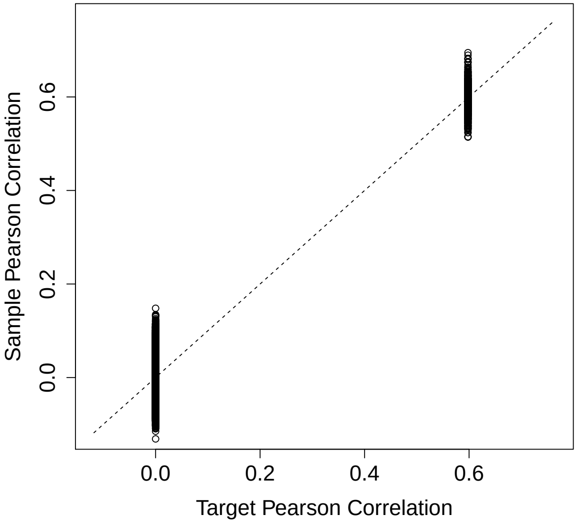
1. Simulate multivariate standard normal vector $\boldsymbol{Z}$ with variance-covariance matrix $\boldsymbol{\Sigma_Z}$ where the elements of $\boldsymbol{\Sigma_Z}$ are either 0, 1, or solutions for δ to the equation corresponding to the pair $(Y_{ij}, Y_{ij'})$.

2. Transform each element of $\boldsymbol{Z}$ to obtain desired marginals:

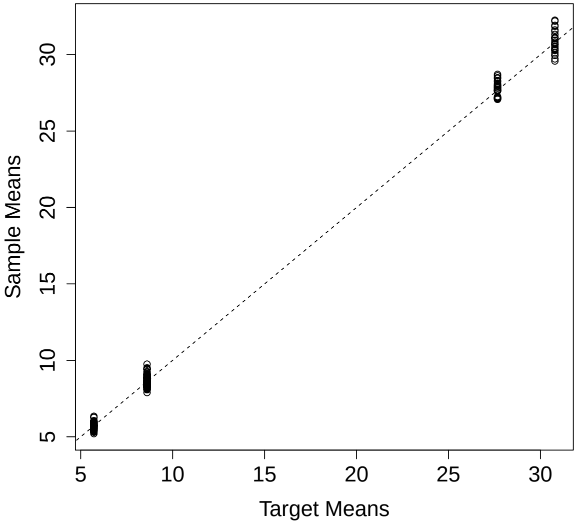
$$Y_{ij} = \widehat{F}_{ij}^{-1}\{\Phi(Z_{ij})\}$$

This process yields one simulated data set. Repeat 1000 times.

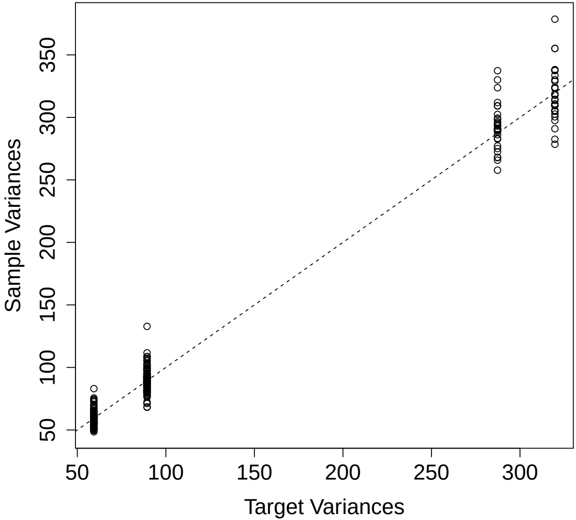
Simulated Seizure Data Results



Simulated Seizure Data Results

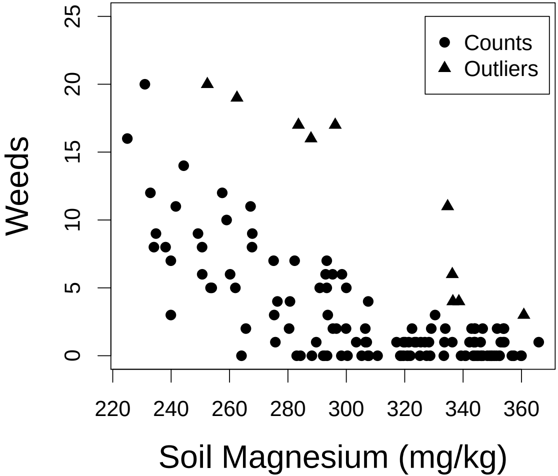


Simulated Seizure Data Results



Outline

Weed Data


Weeds
Soil Magnesium (mg/kg)
Counts
Outliers
0
5
10
15
20
25
220
240
260
280
300
320
340
360

Marginal Model

Negative binomial hurdle model is a Bernoulli mixture of a point mass at 0 and a negative binomial, left-truncated at 1.

$$P(Y = y) = \begin{cases} \pi, & y = 0 \\ (1-\pi) \cdot \dfrac{\dfrac{\Gamma(\theta + y)}{\Gamma(\theta)\Gamma(y+1)} \left(\dfrac{\theta}{\theta+\mu}\right)^{\theta} \left(\dfrac{\mu}{\theta+\mu}\right)^{y}}{1 - \left(\dfrac{\theta}{\theta+\mu}\right)^{\theta}}, & y \geq 1 \end{cases}$$

Model π and negative binomial mean μ as functions of covariate, $x =$ soil magnesium.

Negative Binomial Hurdle CDF

The CDF for Y_i is then

$$F_i(y) = \pi_i + \frac{1-\pi_i}{1-g_i(0|\mu_i,\theta)}\left\{G_i(y|\mu_i,\theta) - g_i(0|\mu_i,\theta)\right\}$$

for $y \geq 0$, where $G_i(\cdot|\mu_i,\theta)$ and $g_i(\cdot|\mu_i,\theta)$ are the negative binomial CDF and PDF with

$$\log(\mu_i) = \beta_0 + \beta_1 x_i \ ,$$

and

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 x_i \ .$$

Negative Binomial Hurdle CDF

The CDF for Y_i is then

$$F_i(y) = \pi_i + \frac{1-\pi_i}{1-g_i(0|\mu_i,\theta)}\left\{G_i(y|\mu_i,\theta) - g_i(0|\mu_i,\theta)\right\}$$

for $y \geq 0$, where $G_i(\cdot|\mu_i,\theta)$ and $g_i(\cdot|\mu_i,\theta)$ are the negative binomial CDF and PDF with

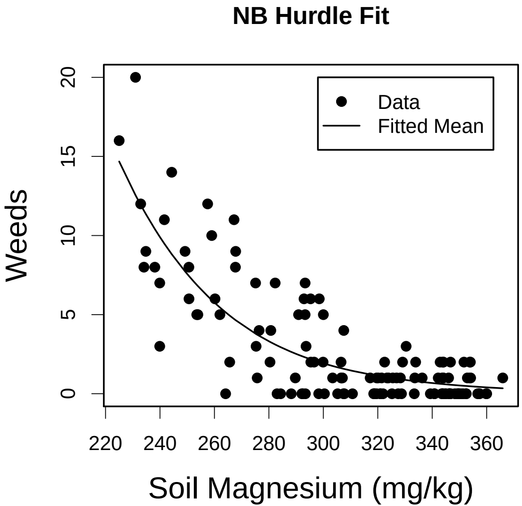
$$\log(\mu_i) = \beta_0 + \beta_1 x_i \ ,$$

and

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 x_i \ .$$

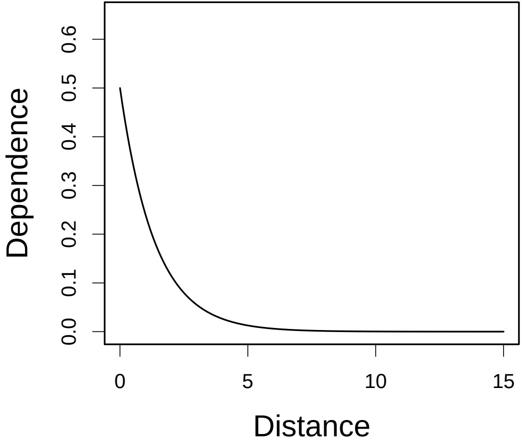
Plug in estimates of $\beta_0, \beta_1, \gamma_0, \gamma_1$, and overdispersion parameter θ to obtain target marginal CDFs.

Weed Data With Fitted Means



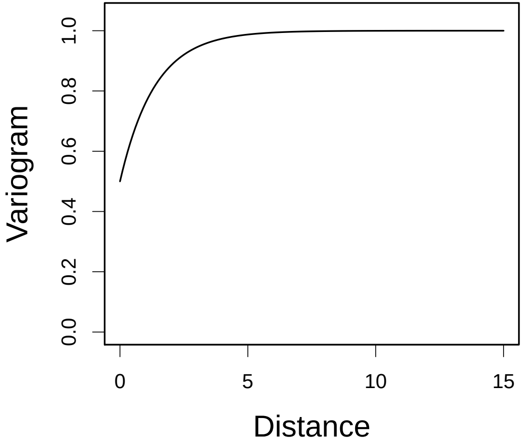
The Principle of Spatial Dependence

Dependence between observations is higher when they are close together.



Variogram Characterizes Spatial Dependent

var($Y_i - Y_j$) is small if Y_i and Y_j are dependent.



Stationarity

A typical spatial data set represents a single incomplete sample of size $N = 1$ from a spatial random process.

Stationarity

A typical spatial data set represents a single incomplete sample of size $N = 1$ from a spatial random process.

To make inference feasible, we assume *stationarity*, i.e. $E(Y_i)$=$E(Y_j)$ and $\text{var}(Y_i - Y_j) = 2\gamma(\boldsymbol{h}_{ij})$, where $\boldsymbol{h}_{ij}$ is the vector between locations of Y_i and Y_j, and $\gamma(\cdot)$ is called the *semivariogram*.

Stationarity

A typical spatial data set represents a single incomplete sample of size $N = 1$ from a spatial random process.

To make inference feasible, we assume *stationarity*, i.e. $E(Y_i)$=$E(Y_j)$ and $\text{var}(Y_i - Y_j) = 2\gamma(\boldsymbol{h}_{ij})$, where $\boldsymbol{h}_{ij}$ is the vector between locations of Y_i and Y_j, and $\gamma(\cdot)$ is called the *semivariogram*.

Weed counts are not stationary: means differ, and larger means are associated with larger variance.

Stationarity

A typical spatial data set represents a single incomplete sample of size $N = 1$ from a spatial random process.

To make inference feasible, we assume *stationarity*, i.e. $E(Y_i)$=$E(Y_j)$ and var$(Y_i - Y_j) = 2\gamma(\boldsymbol{h}_{ij})$, where $\boldsymbol{h}_{ij}$ is the vector between locations of Y_i and Y_j, and $\gamma(\cdot)$ is called the *semivariogram*.

Weed counts are not stationary: means differ, and larger means are associated with larger variance.

Stationarity assumption is more reasonable for ranks than counts.

Ranking Spatial Data

Estimator of ρ_S uses sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, but spatial sample has no replication.

Ranking Spatial Data

Estimator of ρ_S uses sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, but spatial sample has no replication.

Kruskal (1958): Population analog of rank $r(Y_i)$ is $F(Y_i)$.

Ranking Spatial Data

Estimator of ρ_S uses sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, but spatial sample has no replication.

Kruskal (1958): Population analog of rank $r(Y_i)$ is $F(Y_i)$.

For each Y_i, we can estimate its CDF F_i by plugging in point estimates of the parameters.

Ranking Spatial Data

Estimator of ρ_S uses sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, but spatial sample has no replication.

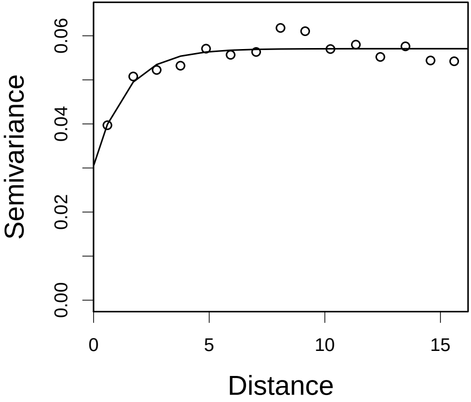
Kruskal (1958): Population analog of rank $r(Y_i)$ is $F(Y_i)$.

For each Y_i, we can estimate its CDF F_i by plugging in point estimates of the parameters.

If Y_i is unusually large (or small), given its estimated distribution, $\hat{F}_i(Y_i)$ will also be unusually large (or small), but $\hat{F}_1(Y_1), \ldots, \hat{F}_n(Y_n)$ will all be on the same scale.

Estimating Spatial Dependence

Fit a parametric semivariogram model to the “ranked” spatial counts.



For Y_i and Y_j separated by a distance of h_{ij},

$$\frac{1}{2}\widehat{\text{var}}[F_i(Y_i) - F_j(Y_j)] =$$

$$0.03 + 0.027\left(1 - e^{-h_{ij}/1.36}\right)$$

$$\Rightarrow \hat{\rho}_{RS}(Y_i, Y_j) = 0.47e^{-h_{ij}/1.36}$$

Calculating $\boldsymbol{\Sigma}_Z$

1. For each pair i, j, obtain

$$\hat{\rho}_S(Y_i, Y_j) = \left\{ \left[1 - \sum_{r=0}^{\infty} \hat{f}_i(r)^3 \right] \left[1 - \sum_{s=0}^{\infty} \hat{f}_j(s)^3 \right] \right\}^{1/2} \cdot \hat{\rho}_{RS}(Y_i, Y_j),$$

where $\hat{f}_i$ and $\hat{f}_j$ are the estimated PMFs of Y_i and Y_j.

Calculating $\boldsymbol{\Sigma}_Z$

1. For each pair i, j, obtain

$$\hat{\rho}_S(Y_i, Y_j) = \left\{ \left[1 - \sum_{r=0}^{\infty} \hat{f}_i(r)^3 \right] \left[1 - \sum_{s=0}^{\infty} \hat{f}_j(s)^3 \right] \right\}^{1/2} \cdot \hat{\rho}_{RS}(Y_i, Y_j),$$

where $\hat{f}_i$ and $\hat{f}_j$ are the estimated PMFs of Y_i and Y_j.

2. Then numerically solve for $\delta = \rho(Z_i, Z_j)$:

$$\begin{aligned}
\hat{\rho}_S(Y_i, Y_j) = 3 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \hat{f}_i(r) \hat{f}_j(s) \big(& \Phi_\delta\{\Phi^{-1}[\hat{F}_i(r-1)], \Phi^{-1}[\hat{F}_j(s-1)]\} \\
& + \Phi_\delta\{\Phi^{-1}[1 - \hat{F}_i(r)], \Phi^{-1}[1 - \hat{F}_j(s)]\} \\
& - \Phi_{-\delta}\{\Phi^{-1}[\hat{F}_i(r-1)], \Phi^{-1}[1 - \hat{F}_j(s)]\} \\
& - \Phi_{-\delta}\{\Phi^{-1}[1 - \hat{F}_i(r)], \Phi^{-1}[\hat{F}_j(s-1)]\} \big).
\end{aligned}$$

Apply Algorithm

Retain locations and covariate values from data set.

1. Simulate a multivariate standard normal vector $\boldsymbol{Z}$ with correlation matrix $\boldsymbol{\Sigma_Z}$.

Apply Algorithm

Retain locations and covariate values from data set.

1. Simulate a multivariate standard normal vector $\boldsymbol{Z}$ with correlation matrix $\boldsymbol{\Sigma_Z}$.
2. Set $Y_i = \hat{F}_i^{-1}\{\Phi(Z_i)\}$.

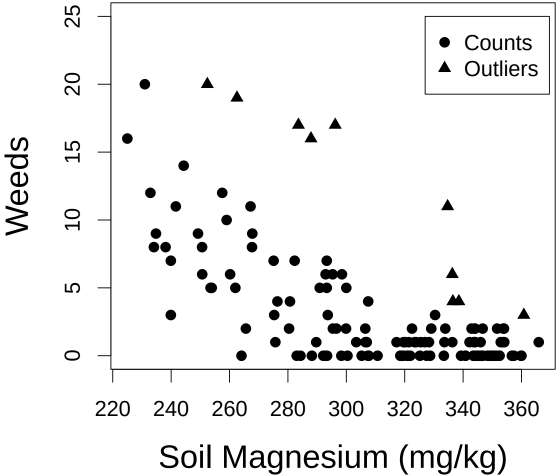
Apply Algorithm

Retain locations and covariate values from data set.

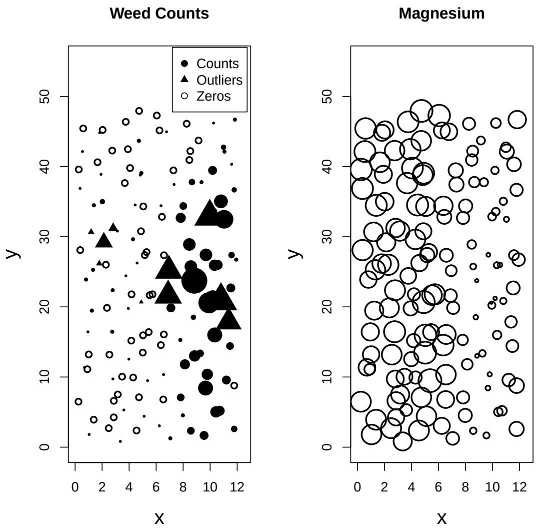
1. Simulate a multivariate standard normal vector $\boldsymbol{Z}$ with correlation matrix $\boldsymbol{\Sigma_Z}$.
2. Set $Y_i = \hat{F}_i^{-1}\{\Phi(Z_i)\}$.

Repeat 1000 times to obtain 1000 data sets.

Two Outlier Processes



Outliers Localized



Empirical Observations About Outliers

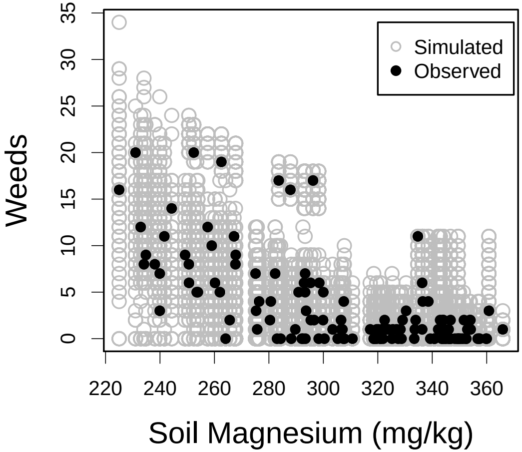
- Outliers occur in the region between $y = 17$ and $y = 33$ meters.
- Outliers associated with mg between 250 and 300 are between 12.9 and 14.9 larger than target means, whereas outliers associated with mg above 330 are between 2.6 and 10.3 larger.

Augmenting the Simulated Data with Outliers

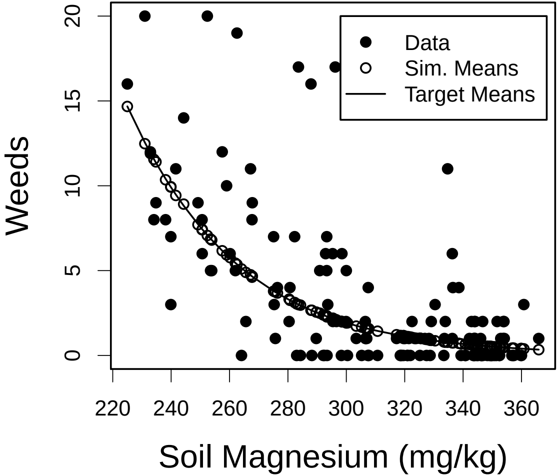
For each of the 1000 simulated data sets,

- Randomly select 4 to 6 locations with y-coordinates between 17 and 33 and mg between 250 and 300.
- Set these counts equal to the integer part of target mean plus a random uniform on $(12, 15)$.
- Randomly select another 4 to 6 points with y-coordinates between 17 and 33 and mg exceeding 330.
- Set these to the integer part of target means plus a random uniform on $(2, 11)$.

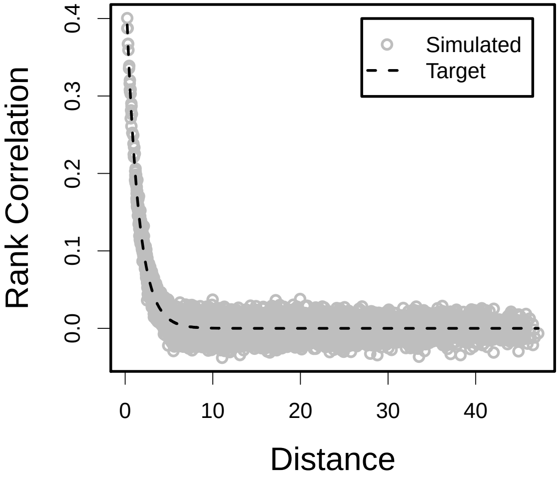
Simulated Data vs. Observed Data



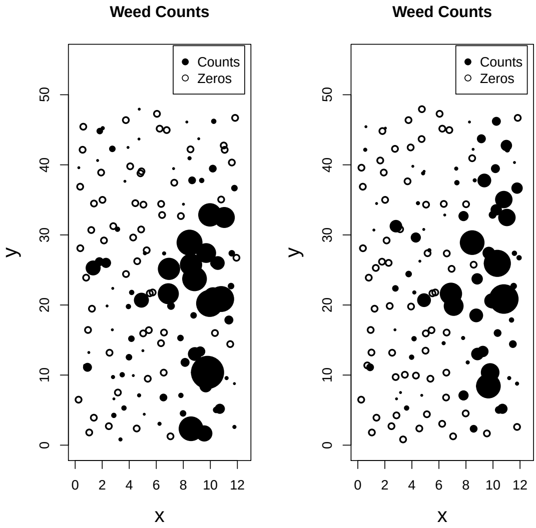
Simulated Data vs. Observed Data



Simulated Data vs. Observed Data



A Couple of Simulated Maps



References

- S. Heijting, W. Van Der Werf, A. Stein, and M.J. Kropff (2007), Are weed patches stable in location? Application of an explicitly two-dimensional methodology, *Weed Research* 47 (5), pp. 381-395. DOI: 10.1111/j.1365-3180.2007.00580.x
- W.H. Kruskal (1958), Ordinal measures of association, *Journal of the American Statistical Association* 53, pp. 814–861.
- L. Madsen and D. Birkes (2013), Simulating dependent discrete data, *Journal of Computational and Graphical Statistics*, 83(4), pp. 677–691.
- J. Nešlehová (2007), On rank correlation measures for non-continuous random variables, *Journal of Multivariate Analysis* 98, pp. 544–567.