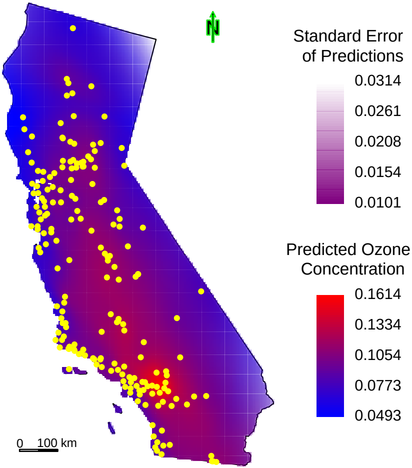
Introduction to Spatial Statistics

Jay Ver Hoef

National Marine Mammal Lab
NOAA Fisheries
International Arctic Research Center
Fairbanks, Alaska, USA

Outline

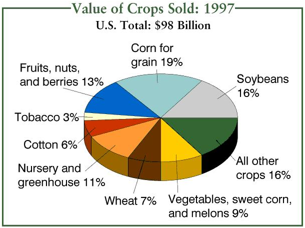
- Introduction
- Autocorrelation
- Types of Spatial Data
- Prediction
- Regression
- Design of Experiments
- Sampling
- Summary

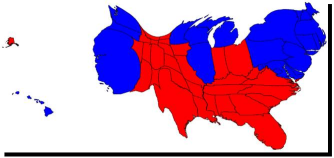


What are Statistics?

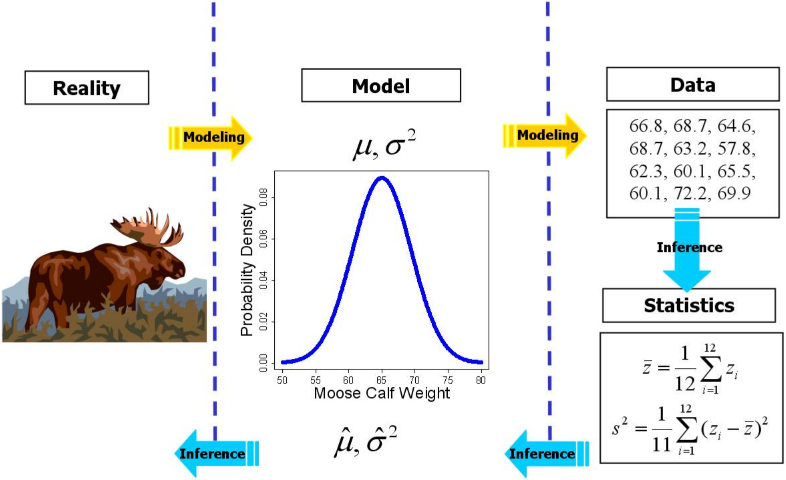
$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

A statistic is a function of data





Statistical Models and Inference

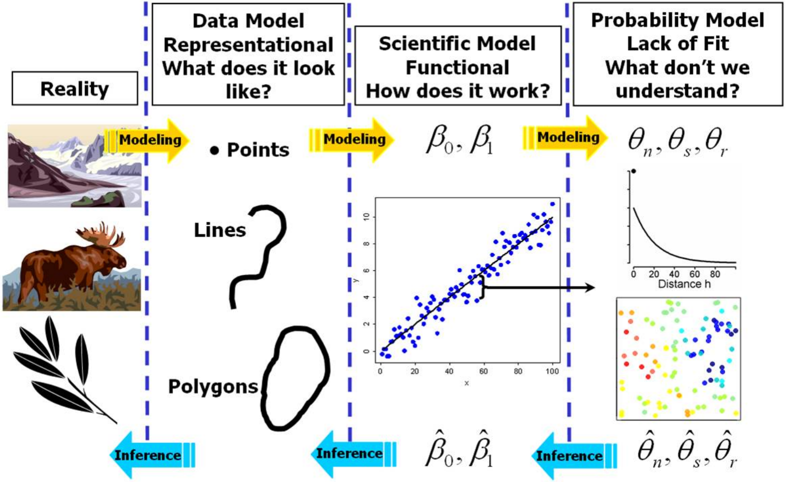


What is a Model?

- What does it look like?
- Structural

- How does it work?
- Functional

Spatial Statistical Models and Inference



Spatial Linear Model

$$z_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + \ldots + \epsilon_i$$

$$z_{j,k,i} = \beta_0 + \tau_j + \delta_k + (\tau\delta)_{j,k} + \ldots + \epsilon_{j,k,i}$$

$$\begin{pmatrix} \mathbf{z}_{\text{observed}} \\ \mathbf{z}_{\text{unobserved}} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

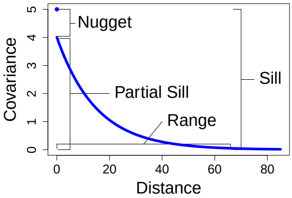
- Point Prediction
- Block Prediction
- Sampling

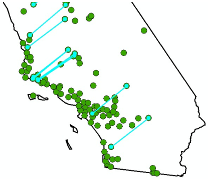
- Regression
- Design of Experiments

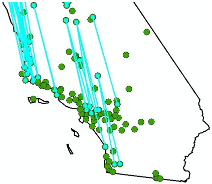
Covariance Matrix in Spatial Linear Models

Reduce the number of parameters in the probability model by using spatial relationships

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{pmatrix}$$

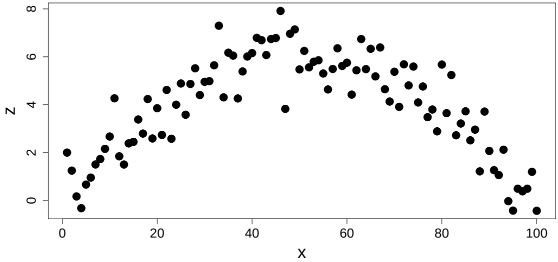






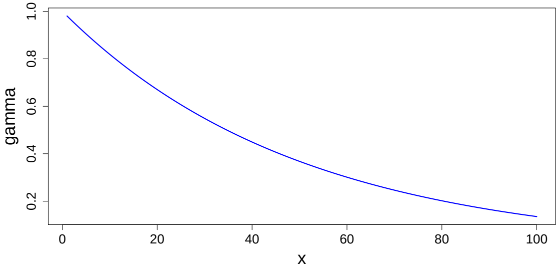
The Many Faces of “Autocorrelation”: DATA

```
x <- 1:100
z <- 6 - ((x - 50)/20)^2 + rnorm(100)
plot(x, z, pch = 19, cex = 2, cex.lab = 2, cex.axis = 1.5)
```



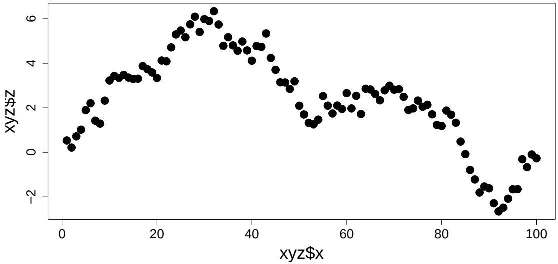
The Many Faces of “Autocorrelation”: MODEL

```
x <- 1:100
gamma <- exp(-x/50)
plot(x, gamma, type = "l", lwd = 2, cex.lab = 2, cex.axis = 1.5, col = "blue")
```



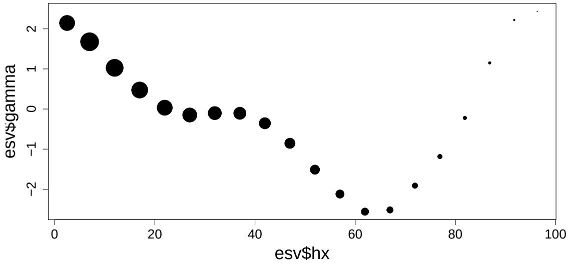
The Many Faces of “Autocorrelation”: PROCESS

```
set.seed(4)
x <- 1:100
y <- rep(1, times = 100)
xyz <- geoStatSim(x, y, range = 100, nugget = 0.01, parsil = 6)
plot(xyz$x, xyz$z, pch = 19, cex = 2, cex.lab = 2, cex.axis = 1.5)
```



The Many Faces of "Autocorrelation": STATISTIC

```
spDF <- SpatialPointsDataFrame(cbind(xyz$x, xyz$y), data.frame(z = xyz$z))
esv <- empSemivariogram(spDF, "z", EmpVarMeth = "CovMean")
plot(esv$hx, esv$gamma, pch = 19, cex = esv$np/100, cex.lab = 2, cex.axis = 1.5)
```



Estimation Versus Prediction

$$E(\mathbf{z}, Z_0) = \mathbf{1}\mu, \quad \mathrm{cov}(\mathbf{z}, Z_0) = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{c} \\ \mathbf{c}' & \sigma_0^2 \end{pmatrix}$$

data mean = $\mathbf{1}'\mathbf{z}/n$

variance as estimator of μ = $\mathbf{1}'\mathbf{\Sigma}\mathbf{1}/n^2$

variance as predictor = $E(\mathbf{1}'\mathbf{z}/n - Z_0)^2 = \mathbf{1}'\mathbf{\Sigma}\mathbf{1}/n^2 - 2\mathbf{1}'\mathbf{c}/n + \sigma_0^2$

```
# Independence
SigmaInd <- diag(6)
# variance of mean estimator for first 5
sum(SigmaInd[1:5, 1:5])/5^2


## [1] 0.2


# variance of first 5 to predict the 6th
sum(SigmaInd[1:5, 1:5])/5^2 - 2 * sum(SigmaInd[6, 1:5])/5 + SigmaInd[6, 6]


## [1] 1.2
```

Estimation Versus Prediction

$$E(\mathbf{z}, Z_0) = \mathbf{1}\mu, \quad \mathrm{cov}(\mathbf{z}, Z_0) = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{c} \\ \mathbf{c}' & \sigma_0^2 \end{pmatrix}$$

data mean = $\mathbf{1}'\mathbf{z}/n$

variance as estimator of μ = $\mathbf{1}'\mathbf{\Sigma}\mathbf{1}/n^2$

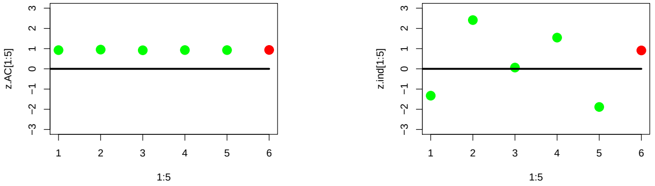
variance as predictor = $E(\mathbf{1}'\mathbf{z}/n - Z_0)^2 = \mathbf{1}'\mathbf{\Sigma}\mathbf{1}/n^2 - 2\mathbf{1}'\mathbf{c}/n + \sigma_0^2$

```
# lots of autocorrelation
SigmaAC <- matrix(0.9999, nrow = 6, ncol = 6)
diag(SigmaAC) <- 1
# variance of mean estimator for first 5
sum(SigmaAC[1:5, 1:5])/5^2

## [1] 0.9999

# variance of first 5 to predict the 6th
sum(SigmaAC[1:5, 1:5])/5^2 - 2 * sum(SigmaAC[6, 1:5])/5 + SigmaAC[6, 6]

## [1] 0.00012
```

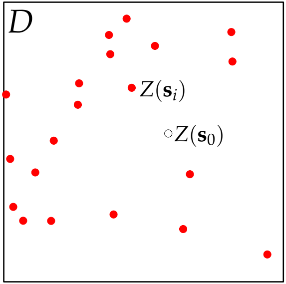
Estimation Versus Prediction

$$\begin{pmatrix} 1.0000 & 0.9999 & 0.9999 & 0.9999 & 0.9999 & 0.9999 \\ 0.9999 & 1.0000 & 0.9999 & 0.9999 & 0.9999 & 0.9999 \\ 0.9999 & 0.9999 & 1.0000 & 0.9999 & 0.9999 & 0.9999 \\ 0.9999 & 0.9999 & 0.9999 & 1.0000 & 0.9999 & 0.9999 \\ 0.9999 & 0.9999 & 0.9999 & 0.9999 & 1.0000 & 0.9999 \\ 0.9999 & 0.9999 & 0.9999 & 0.9999 & 0.9999 & 1.0000 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

```
set.seed(18)
z.AC <- t(chol(SigmaAC)) %*% rnorm(6)
z.ind <- t(chol(SigmaInd)) %*% rnorm(6)
plot(1:5, z.AC[1:5], ylim = c(-3, 3), xlim = c(1, 6), pch = 19, cex = 2, col = "green")
points(6, z.AC[6], pch = 19, cex = 2, col = "red")
lines(c(0, 6), c(0, 0), lwd = 3)
plot(1:5, z.ind[1:5], ylim = c(-3, 3), xlim = c(1, 6), pch = 19, cex = 2, col = "green")
points(6, z.ind[6], pch = 19, cex = 2, col = "red")
lines(c(0, 6), c(0, 0), lwd = 3, pch = 19, cex = 2)
```

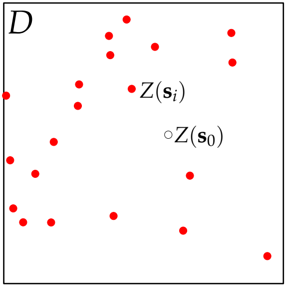


Notation



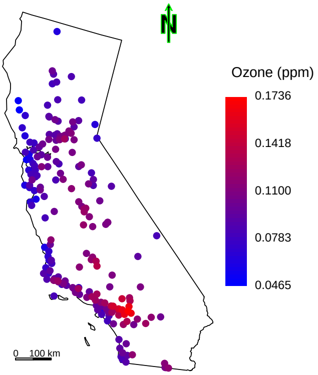
- D is the spatial region or area of interest
- $\mathbf{s}$ contains the spatial coordinates
- Z is the value located at the spatial coordinates

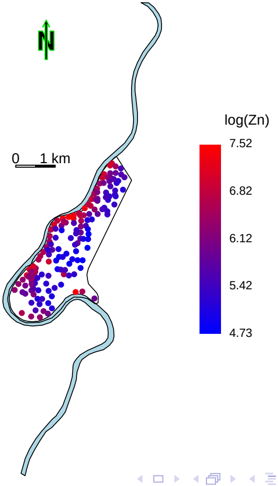
Types of Spatial Data



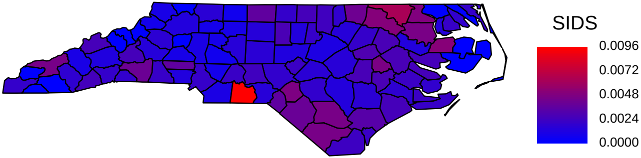
- $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$
- Geostatistical Data: Z random, D fixed, continuous, infinite
- Lattice/Aerial Data: Z random, D fixed, finite, (ir)regular grid
- Point Pattern Data: $Z \equiv 1$, D random, finite

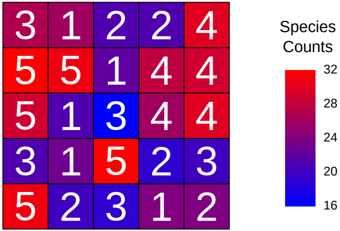
Examples of Geostatistical Data





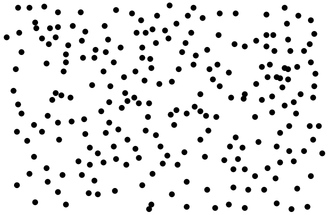
Examples of Lattice/Aerial Data



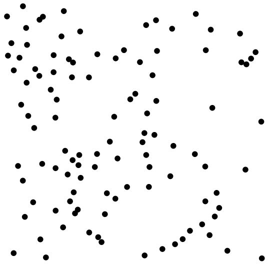


Examples of Point Pattern Data

Anemones



Ponderosa



Estimation

minimize for θ

$$-2\ell(\boldsymbol{\theta}, \mathbf{z}) \propto \log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \mathbf{r}'_{\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1}\mathbf{r}_{\boldsymbol{\theta}}$$

or

$$-2\ell_{\mathrm{REML}}(\boldsymbol{\theta}, \mathbf{z}) \propto \log|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| + \mathbf{r}'_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{r}_{\boldsymbol{\theta}} + \log|\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}|$$

where

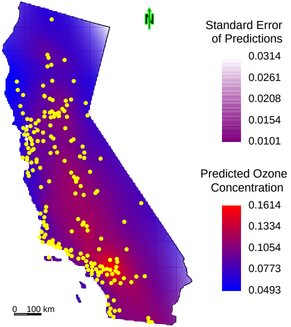
$$\mathbf{r}_{\boldsymbol{\theta}} = \mathbf{z} - \mathbf{X}\hat{\beta}_{\boldsymbol{\theta}}$$

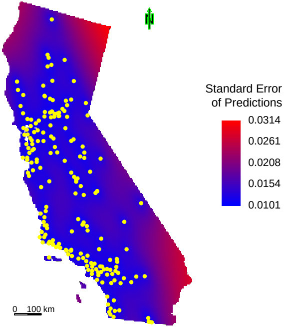
and

$$\hat{\beta}_{\boldsymbol{\theta}} = (\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{z}$$

Prediction

$$\begin{pmatrix} \mathbf{z}_{\text{observed}} \\ \mathbf{z}_{\text{unobserved}} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \ \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$





Spatial Regression

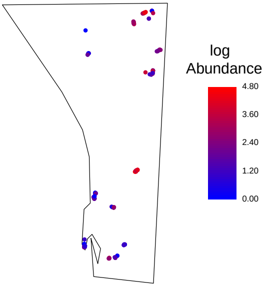
- Whiptail Lizard
- 148 locations in Southern California
- Measured the average number caught in traps over 80-90 trapping events in one year
- Data log-transformed, one outlier removed

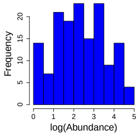
Whiptail Lizard Data

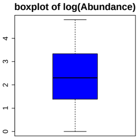
$$\begin{pmatrix} \mathbf{z}_{\text{observed}} \\ \mathbf{z}_{\text{unobserved}} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \ \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

- Ant Abundance
- Percent Sandy Soil

- Matern Model
- Anisotropy







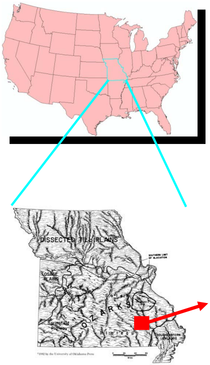
Fitted Model for Whiptail Lizard Data

$$\begin{pmatrix} \mathbf{z}_{\text{observed}} \\ \mathbf{z}_{\text{unobserved}} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \;\; \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

Effect	Est	Std Error	t-value	df	Pr(t:H$_0$)
Intercept	0.716	0.574	146	1.25	0.2139
Ant Abund	0.252	0.107	146	2.36	0.0195
Sandy Soil	0.764	0.249	146	3.07	0.0026

Component	Parameter	Estimate
nugget	nugget	0.598
besselK	parsil	1.027
besselK	range	160313
besselK	minorp	0.042
besselK	rotate	18.5
besselK	extrap	0.539

Glades in the Ozarks



Simulated Spatial Experimental Design

32	26	24	24	24
26	25	22	22	23
23	21	21	20	24
26	23	26	22	25
25	23	24	24	27

Treatment	Effect
1	0
2	–3
3	–5
4	+6
5	+6

Add Trts

3	1	2	2	4
5	5	1	4	4
5	1	3	4	4
3	1	5	2	3
5	2	3	1	2

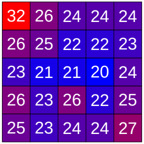
Estimate

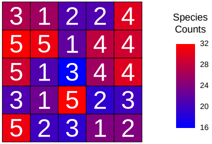
Contrast	*True Value*
$c_1 = (\tau_2 + \tau_3)/2 - \tau_1$	-4.00
$c_2 = (\tau_4 + \tau_5)/2 - \tau_1$	6.00
$c_3 = (\tau_4 + \tau_5)/2 - (\tau_2 + \tau_3)/2$	10.00
$c_4 = (\tau_2 - \tau_3)$	2.00
$c_5 = (\tau_4 - \tau_5)$	0.00

Simulated Spatial Experimental Design

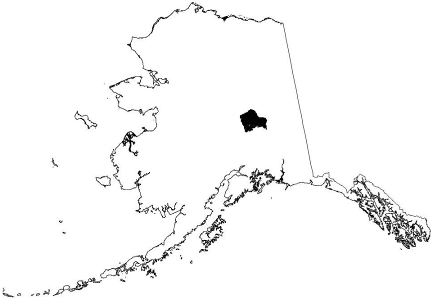
$$\begin{pmatrix} \mathbf{z}_{\text{observed}} \\ \mathbf{z}_{\text{unobserved}} \end{pmatrix} = \mathbf{X}\beta + \boldsymbol{\epsilon},\ \ \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$$

True Value	Ind Est	Ind SE	Sp Est	Sp SE
-4	-2.4	1.29	-2.95	0.87
6	6.6	1.29	6.81	1.05
10	9.0	1.05	9.77	0.84
2	0.4	1.49	0.53	1.07
0	-2.4	1.49	-1.94	1.68
	nugget: 5.56		nugget: 0.00	
			partial sill: 13.55	
			range: 9.36	





Spatial Sampling



- Moose Survey
- South of Fairbanks
- $\sim$ 4500 mi^2

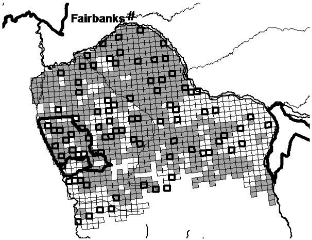
Total Area

SRS

- $\hat{\tau} = 11535$
- $\mathrm{se}(\hat{\tau}) = 985$

FPBK

- $\hat{\tau} = 11327$
- $\mathrm{se}(\hat{\tau}) = 978$



Small Area

SRS(n=17)

- $\hat{\tau} = 1535$
- $\mathrm{se}(\hat{\tau}) = 227$

FPBK

- $\hat{\tau} = 1437$
- $\mathrm{se}(\hat{\tau}) = 153$