1999

Modeling of catastrophic events with applications to insurance

Mary Noga
University of Northern Iowa

Modeling of Catastrophic Events with Applications to Insurance

By Mary Noga

April 11, 1999

Senior Thesis

Advisor Dr. Syed Kirmani

Signatures for completion of senior thesis.

Modeling of Catastrophic Events

With Applications to Insurance

Dr. Syed Kirmani, Mathematics Department

Tim Lindquist, Chair-Presidential Scholars Board

Tornadoes affect every are in the continental United States. There are great differences in the number and amount of damages caused by tornadoes in different regions. The mid-west is nicknamed tornado alley, because of the large number of tornadoes in this region each year. A tornado is commonly described as a "rapidly rotating, slender, funnel-shaped cloud.

The National Climatic Data Center (NCDC) has data on tornadoes over the past four years. This data includes the state, date, and damage done by each tornado. For the purpose of this study a data-set including all tornadoes causing over ten million in damages was retrieved. Table 1 shows the forty data points found.

The study of tornado damages is important for all insurers as all geographical locations in the United States have tornadoes. The damages done by tornadoes are rare and large. Such values are mathematically defined as extreme values. According to Emil Gumbel one purpose of studying extreme values "is to explain observed extremes arising of samples of given sizes…" (qtd. in Hickman). Many times experience alone will not enable the insurer to accurately insurer against losses. Therefore, it may be appropriate to attempt to model tornado damages. This would better enable the insurer to prepare for future losses.

Table 1

State	Region	Damage	Month	Year
30	5	24	6	1995
30	5	30	6	1995
2	5	150	4	1996
2	5	156	4	1996
12	2	13	4	1996
12	2	100	5	1996
18	2	12	4	1996

21	2	12	5	1996
32	2	40	7	1996
6	5	14	11	1997
16	2	90	7	1997
16	2	30	7	1997
29	5	45	3	1997
30	5	12	1	1997
30	5	40	5	1997
30	5	70	5	1997
30	5	15	5	1997
1	5	200	4	1998
6	5	20	2	1998
6	5	175	2	1998
6	5	30	2	1998
6	5	15	2	1998
6	5	30	2	1998
6	5	50	2	1998
7	5	15	3	1998
7	5	15	4	1998
7	5	25	4	1998
7	5	50	4	1998
7	5	38	4	1998
17	2	30	3	1998
17	2	45	3	1998
17	2	120	3	1998
17	2	20	3	1998
17	2	20	3	1998
20	3	34	3	1998
20	3	50	5	1998
23	3	60	5	1998
24	3	20	6	1998
28	2	17	5	1998
29	5	100	4	1998

Exploratory Data Analysis

As in most statistical problems the data is most easily explained graphically. Therefore, several techniques were implemented to graphically display the data before attempting to fit a model.

First a histogram of the damages was prepared. See Appendix A. This histogram has a long tail to the right. This is typical for the lognormal distribution. Additionally, a normal probability plot (NPP) was prepared, Appendix B. If the data was distributed normally, the NPP would appear as a straight line. However, from the histogram it is

assumed the data is not normally distributed. Hence, the NPP is not linear.

To test linear regression parameters, normal data are required. Often a transformation of the data is required before the data can be used in a regression format. It is hypothesized that the natural log of the damages may appear more normally distributed. Therefore, a histogram and NPP are graphed using the natural log of the damages, Appendices C and D. The histogram appears approximately normal, although there is still a slight tail to the left. Additionally, the NPP is approximately linear. Using a pencil most of the points can be covered, exceptions include a few outliers at both ends. Therefore, a regression analysis was preformed on the natural log of the damages.

Regression Analysis

Since the natural log of the damages is the variable of interest it is determined as the response variable. Pearson Correlation Coefficients were prepared for each covariate and the response variable. All of the correlations were less than three tenths indicating only a weak association. This does not give a good outlook for a linear regression model. When a model is fit with all of the covariates the r-squared value is only 0.0496. For a good model this value is close to one. Therefore, this model does not seem reasonable. It is possible a model including a reduced number of covariates would provide a better fit.

Five methods were applied to find this "better" model. These methods were comparing root mean squared error, Mallow's Cp, forward selection, backward elimination, and stepwise. Unfortunately, none of these methods yielded a good model. For example, in forward selection, backward elimination, and the stepwise procedure all of variables

failed the test to be in the model. Therefore, none of the variables significantly explain a linear relationship with the natural log of the damages. Because of the complex and unknown nature of tornadoes, it is not surprising a linear regression does not explain the data set. Therefore, more complex techniques will be applied to fit a non-linear model.

Back to Exploratory Data Analysis

Several other multivariate techniques were used to graph the data. Star plots and faces are shown in Appendices E and F. Each ray (direction) of the star represents a different variable. From this graphical display there seems to be several groups of similar tornadoes. For example, numbers 18, 19, 20, 22, 27, 28, 31, 33, and 34 appear similar. A closer look at the data reveals all of these tornadoes were in region five, state six or seven, and in February, March, or April. The faces are an alternate graphical display of the data. Each facial feature represents a variable. For example, the shape of the face represents the region the tornado took place in. Similar patterns, as found in the stars, should be and are apparent in the faces.

Finally, some analysis was done in the attempt to cluster the data by month or region. Appendices G-K show several graphs. First, simple plots of the damages were done by month and region. It can easily be seen that February through May are the typical months for damaging tornadoes. Hardly any tornadoes took place between August and January. Regions two, three, and five were the only regions to have damaging tornadoes in these years. These represent the Central, Eastern, and Southern regions, respectively. Cluster analysis was

preformed to see if tornadoes in the regions would cluster together. The definitions of the regions were given by the national weather service. There appears to be some clustering by region. For example, tornadoes 8, 9, 39, 40 are all in region five and are clustered together and apart form the others. Similar clustering occurs by month. Again 8, 9, and 39 are in the same month: April. However, tornado number forty is in February.

Hypothesis

Let N represent the number of losses each year in excess of ten million dollars. Also, let Xi, for i=1,2,…,n, represent the dollar amounts of the N losses. The distributions of the Xi's are assumed to be continuous. There are K=4 observation years available. The following hypothesis are examined:

(H1) N and (X1,X2,…) are independent random variables i.e. the frequency and severity of losses are independent and random.

(H2) X1,X2,…are independent identically distributed random variables.

Tests of Hypotheses

(H1) Grouping by frequency and severity and using a Chi-Square Independence Test can check the first hypothesis. Four frequency and four severity groups were created, observed(expected):

Yearly Frequency / Loss	(10-57.5)	(57.5-105)	(105-152.5)	(152.5-200)	Total
0-6	2(1.5)	0(0.25)	0(0.15)	0(0.1)	2
7-12	10(11.25)	3(1.875)	2(1.125)	0(0.75)	15
13-18	0	0	0	0	0
19-24	18(17.25)	2(2.875)	1(1.725)	2(1.15)	23
Total	30	5	3	2	40

Chi-Square Independence Test yields TS=4.143 which is less than the CV= 16.92=chisquare(9 d.f., alpha=.05). So, (H1) is accepted.

(H2) Two non-parametric tests will be used to check the second hypothesis, including: Kendall, and Spearman. Specifically, they test the independence of X1 and X2.

Kendall Tau:

Definition

T=2/[4*(4-1)]*sum(Aij)

Aij=1, (Xj1-Xi1)*(Xj2-Xi2)>0

0, " "=0

-1, " "<0

Calculations

T=0, which is less than any possible critical value so we accept that X1 and X2 are independent under the Kendall test.

Spearman Rho:

Definition

Let **X1**'=(24,150,14,200), **X2**'=(30,12,14,15), **Rk** and **Sk** be the rank vectors of **X1** and **X2**, respectively.

R={sum[(Rk-Rbar)*(Sx-Sbar)]}/

{[sum(Rk-Rbar)]^0.5*[sum(Sk-Sbar)]^0.5

Calculations

The observed value of R is zero, which implies X1 and X2 are independent at any alpha level.

Loss Frequency

Let N=(2,7,8,23) represent the observed values of the loss frequencies. The mean and standard deviation are 10 and 61.5, respectively.

The Poisson distribution is a discrete distribution often used to model frequencies. E(N)=V(N)=q

$$P(N=n)=e^{-q}*q^n/n!$$

The maximum likelihood estimator for q is the mean of the sample, 10.

Additionally, the negative binomial can be used to model such frequencies. E(N)=r(1-p)/p, V(N)=r(1-p)/p^2

$$P(N=n)=\Gamma(r+n)/[\Gamma(r)*n!]*p^r*(1-p)^n$$

The method of moments estimators for r and q are 0.1626 and 1.942, respectively.

The following table compares the observed frequencies and those expected by the fitted distributions.

Yearly Frequency	Observed	Poisson(10)	Neg. Bin (.1626,1.942)
0-6	1	.5206	1.596
7-12	2	2.6457	1.219
13-18	0	.8050	.656
19-24	1	.0286	.305
Total	4	4	3.78

The distributions were tested by the Chi-Square Goodness of Fit test. The test statistic is:

TS={sum[(obs-exp)^2]}/exp

This value is distributed Chi-Square with two degrees of freedom, which yields a critical value of 10.6 with alpha=.05.

TS(Poisson)=34.45,Ts(Neg. Bin.)=2.96. Therefore, the negative binomial is acceptable as a good fit to the data.

Loss Amount

Let Xi represent the loss amounts in excess of ten million. X1, X2, X3… are assumed to be independent and identically distributed. Several techniques are used to model the unknown distribution.

Non-parametric estimation:

Let X(1),X(2),X(3)… be the ordered losses by increasing amounts. Fn(x)=k/n, X(k)<=x<=X(k+1) Levi and Partrat describe a non-parametric procedure for estimating 1-F(x) using Fn(x), In(x), and Sn(x). Dn(1-alpha) is the one minus alpha Kolmogorov-Smirnov statistic (260). For alpha=.05, Dn(.95)=.21503. Let In(x) and Sn(x) be defined as follows:

$$In(x)=\max[1-Fn(x)-Dn(1-\alpha),0]$$

$$Sn(x)=\min[1-Fn(x)+Dn(1-\alpha),1]$$

In(x) and Sn(x) form a 95% confidence region for 1-F(x). This is best shown by graph in Appendix L.

Parametric Estimation:

Three distributions are commonly used to model losses of this magnitude: including Exponential, Pareto, and Lognormal.

Exponential with (B;10)

$f_{B,10}(x)=Be^{-B(x-10)}$

$F_{B,10}(x)=1-e^{-B(x-10)}$

The maximum likelihood Estimate of B is:

$\hat{B}=n/\sum(y_i-10)=.0246$

This corresponds to a 95% confidence interval for B^ as follows:

[B^*chisq(2n,alpha/2)/(2n),B^*chisq(2n,1-alpha/2)/(2n)]

=[.01757,.03278]

Additionally, a confidence region for 1-F(x) can be found:

[exp(-.03278(x-10)),exp(-.01757(x-10))]

This can be graphically displayed as in appendix M.

Pareto with (y;10)

gy,10(x)=(y10^y)/(x^(y+1))

Gy,10(x)=1-(10/x)^y

The maximum likelihood estimator of y is:

Y^=n/sum(ln(yi/10))=.78705

Similar to the exponential, this corresponds to a 95% confidence interval for y^ as follows:

[.56225,1.0487]

Additionally, a confidence region for 1-F(x) can be found:

[(10/x)^.56225,(10/x)^1.0487]

This can be graphically displayed as in appendix N.

Log Normal with (u,q;10)

hu,q,10(x)=1/(sqrt(2piq(x-10))*exp(ln((x-10)-u)^2)/(-2q^2))

Hu,q,10(x)=Io((ln(x-10)-q)/q)

The maximum likelihood estimators of u and q are:

U^=1/n*sum(ln(yi-10))=3.0075

Q^=sqrt(1/n(sum[ln(yi-10)-u^]^2))=1.2654

Here the confidence intervals are not as easily derived and are therefore left for possible consideration, upon inconclusive decisions from the following numeric tests.

Let F(x), G(x), and H(x) be defined as above. In order to decide which of these provide the best fit for the data two numeric tests are preformed.

Chi-Square Goodness of Fit:

Let the test statistic be defined:

Qj=sum[(ni-ei)^2]ei

Also let i=1 imply x in (10,57.5)…i=4 imply x in (152.5,inf). The following table contains the required values for the calculations of Qf, Qg, and Qh.

Xi	ni	ei/f	ei/g	ei/h
(10-57.5)	30	27.21	29.9	29.94
(57.5-105)	5	8.7	3.81	5.61
(105-152.5)	3	2.78	1.6	1.98
(152.5,inf)	2	1.31	4.69	2.47

So, Qf=2.24, Qg=3.14, Qh=0.68. Qf and Qg are distributed chi-square with two degrees of freedom. However, Qh is distributed chi-square with one degree of freedom, because of the extra estimated parameter. This yields critical values of 5.99, 5.99, and 3.84, respectively. Each of the models is acceptable under an alpha-level of 0.05. However, under the goodness of fit method a lower number is preferred. Therefore, model H, log normal, would be selected.

Kolmogorov-Smirnov:

In the K-S test Fn(x) is compared to the fitted values of Fj(x) for each j=F, G, and H. The test statistic is the greatest difference in absolute value. This difference always occurs at an endpoint of an interval. Below are the values necessary for calculation.

Xi	Fn(x)	Ff(x)	Fg(x)	Fh(x)
10	0	0	0	0
57.5	.75	.6802	.7476	.7486
105	.875	.8977	.8429	.8888
152.5	.95	.9673	.8829	.9382
200	1	.9895	.9054	.9616

So, KSf=.0698, KSg=.0946, and KSh=.0384. The critical value of a KS test does not follow any known distribution. Rather for alpha equal to .05 is 1.36/sqrt(n)=.215. Once again all of the models would be acceptable under these criteria. However, the KS test prefers the model

with the smallest test statistic. Therefore, the log normal distribution is selected as best under both KS and goodness of fit criteria.

Principal Component Analysis

The goal of a principal component analysis is to transform a data set into a smaller number of components, which are uncorrelated and explain the data well. For the principal component analysis the following variables were included: region, time, damage, and F-value of the tornado. The state was not used as the numbers were assigned alphabetically; therefore, they had no consistent meaning. The time variable was the month the tornado occurred within the time frame of the data. For example, May 1997 was assigned time twenty-nine.

Usually, PCA is preformed using the covariance matrix; however, the correlation matrix should be used in place of the covariance matrix if the data has vastly different scales (Rencher, 43). Since damages are measured in millions while the other variables are small counting numbers the correlation matrix will be used.

Using S-Plus the following principal components were identified:

	Comp.1	Comp. 2	Comp. 3	Comp. 4
Region	0.654	0.262	0.190	0.684
Time	0.435	0.254	-0.815	-0.286
Damage	-0.141	0.895	0.306	-0.293
F-Value	0.603	-0.258	0.453	-0.604

It is decided that enough principal components will be kept to explain approximately eighty percent of the variability. Appendix O shows the cumulative variability explained for each principal component. Therefore, the first three principal components are retained for further analysis. Appendix P shows the loadings (weight) given to each variable

in the components. Component one has large loadings in the region and F-value variables. This could possibly represent some type or cause of the tornado. Maybe a tornado caused by spring storms in the mid-west is inherently different than a tornado occurring within a hurricane in Florida. This difference could be based on both where a tornado occurred, as well as, the strength of the winds within the tornado. Component two has a heavy loading in damage and approximately equal loadings for the other variables. Additionally, component three has larger loadings for the time and F-value variables. Finally, Appendix Q shows a plot of the four principal components.

Discrimination Analysis

A discrimination analysis was preformed with three groups: Low (x,30), Medium (30,x,100), and High (x>100) damages. The region, time, and F-value variables were considered. The linear combination **ei%*%xj** is the ith sample discriminant function. The first and second **ei**'s are listed in the following table.

E1	E2
-1.1704	-0.1401
0.0093	-0.0742
-0.1207	0.5254

The first discrimination function gives much weight to the region where the tornado took place. From Appendix H, it is noted all of the most damaging tornadoes took place in region 5, the south. Therefore, it is not surprising the region is an important factor in discriminating damage groups. One the other hand, the second discrimination function gives much weight to the F-value. The F-value is a measurement of wind speed. It would be expect that with stronger winds damage would increase.

The two dimensional plot of the Fisher's discriminant functions, Appendix R, shows the low damage tornadoes are most easily discriminated. There appears to be little separation between medium and high damage tornadoes. Perhaps this is due to a lack of a better location variable. Such a variable would take population into consideration.

Conclusion

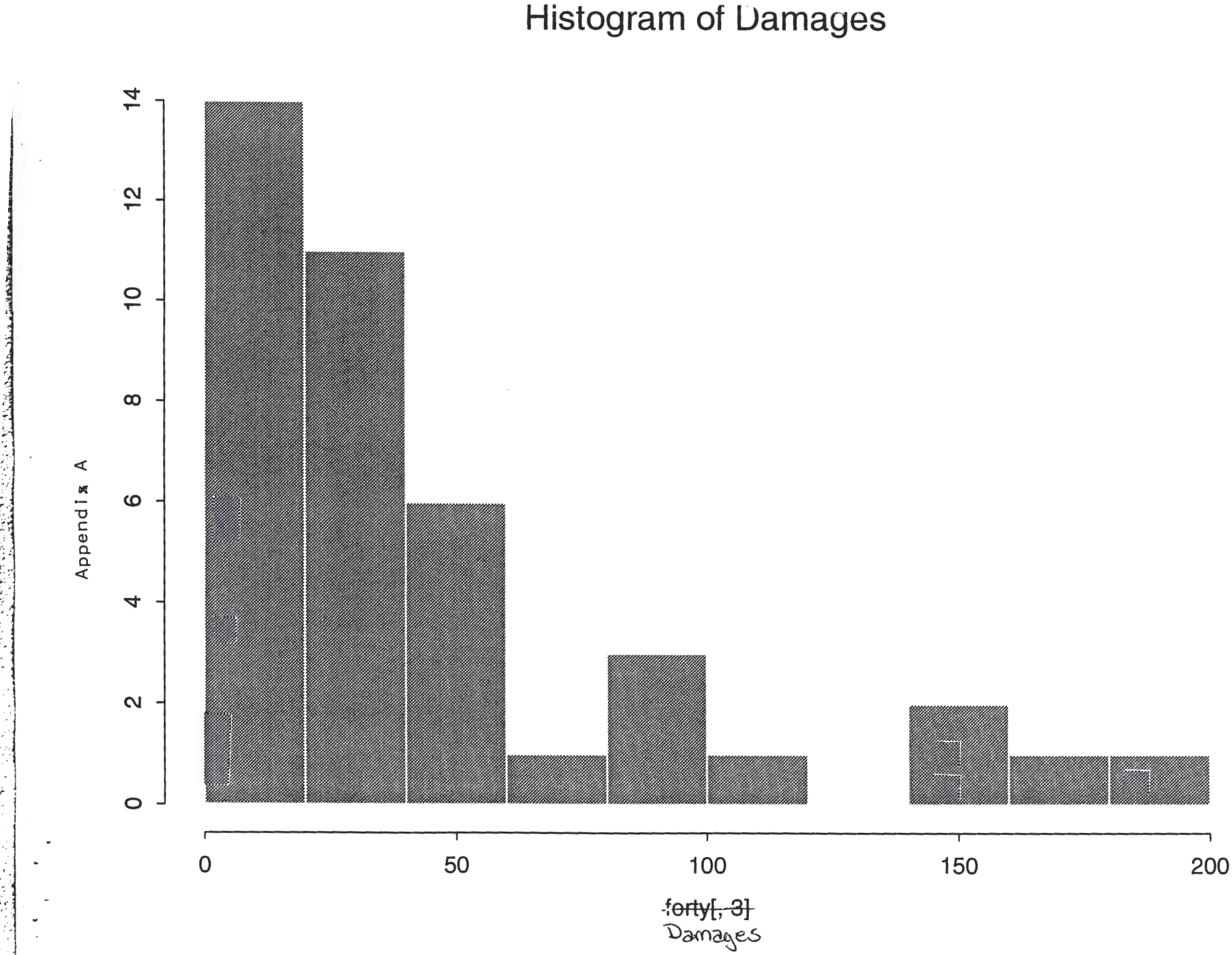
Many techniques have been applied to the tornado data to attempt to model the damages. Unfortunately, the linear regression was unsuccessful due to no covariate being significant. That is often the case with naturally occurring data. However, the modeling using the negative binomial and lognormal distribution proved successful. The principal components transformed the data into three uncorrelated components which explained the data well. Meanings for these components were not completely clear. Finally, the discrimination function was moderately successful at separating the tornadoes by damage amount.

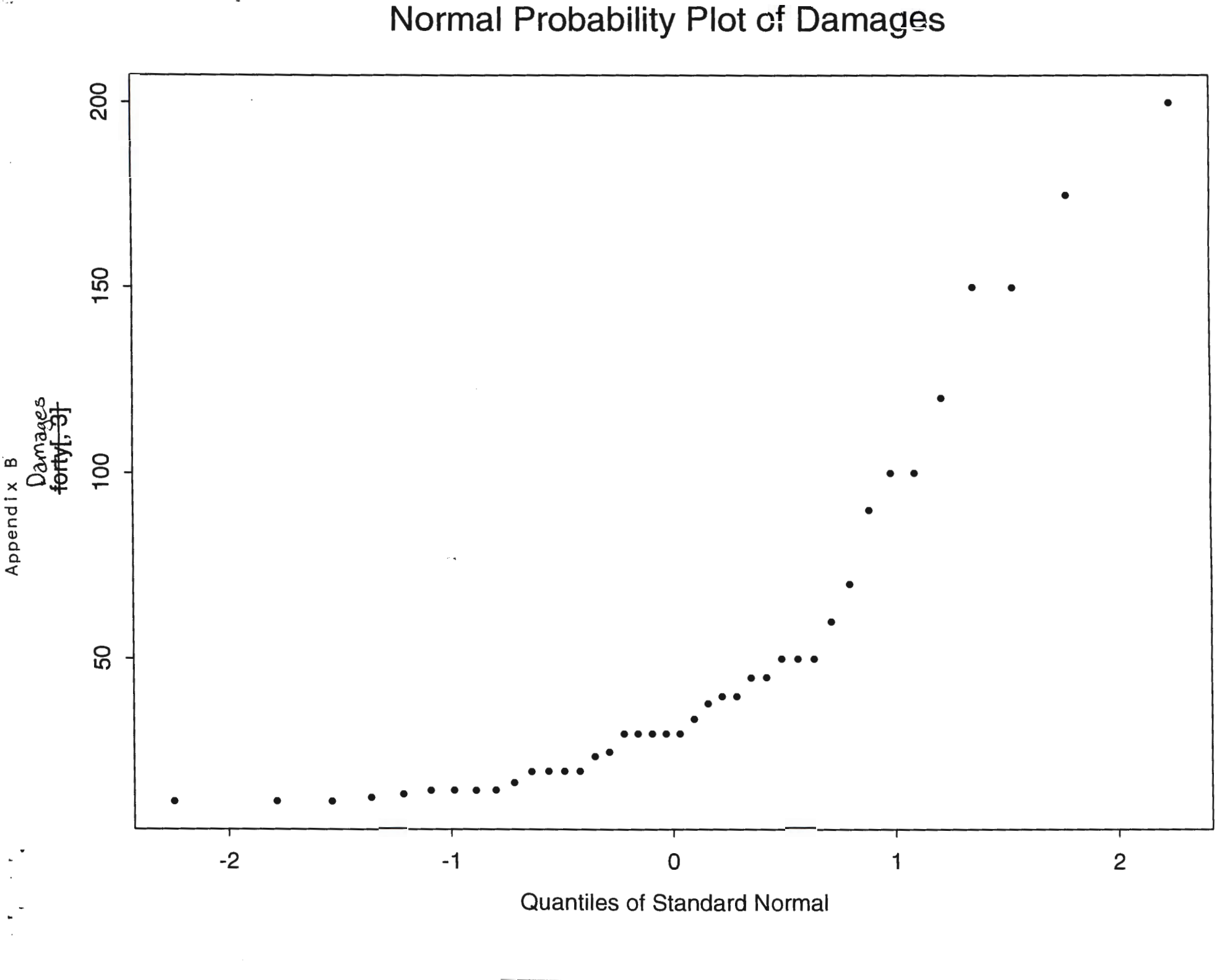
Works Cited

Hickman, Kathyrn and Steven Wall. "Statistics and Extreme Values." The Pentagon 58 (1998): 37-41.

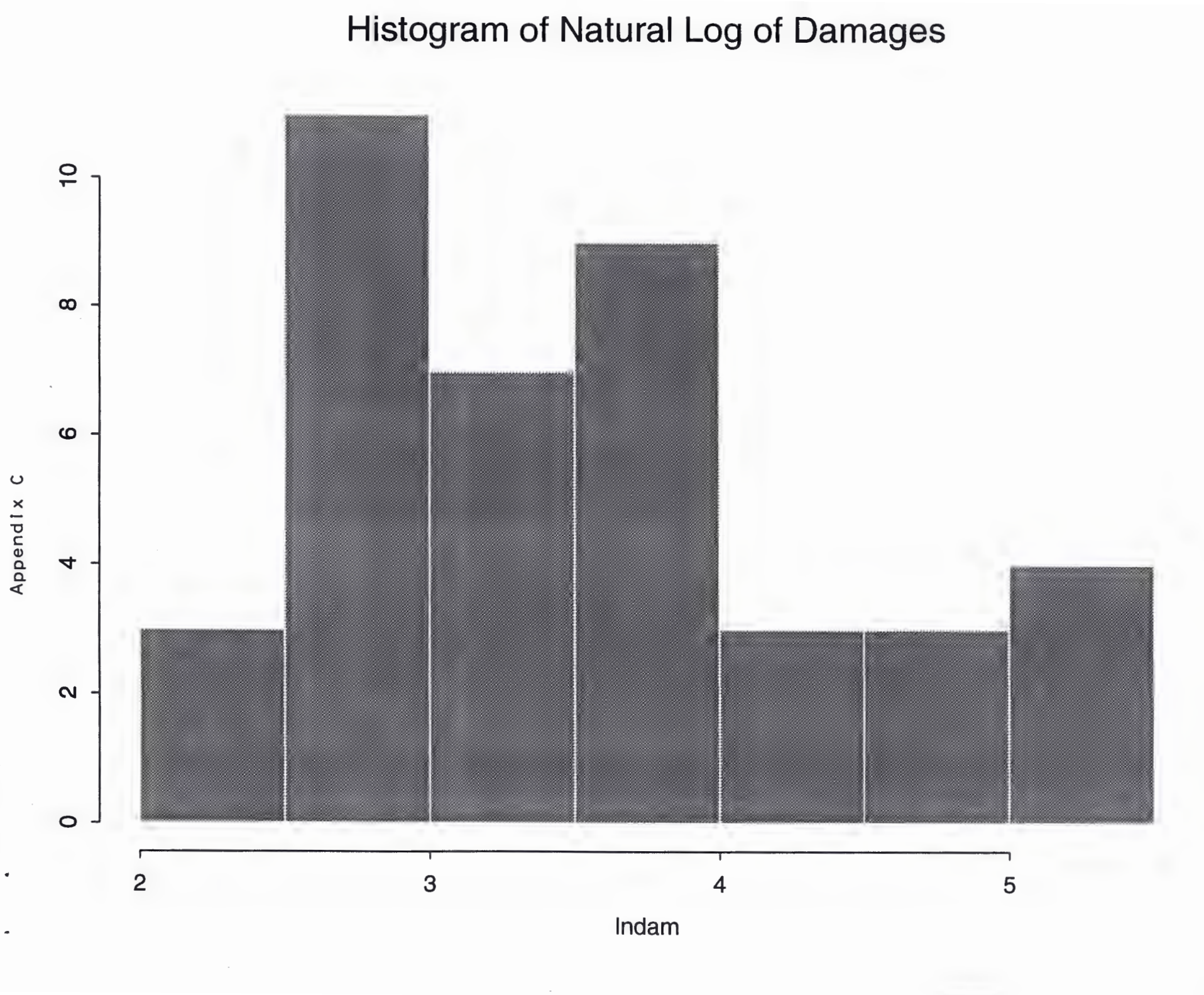
Levi, Charles and Christian Partrat. "Statistical Analysis of Natural Events in the United States." ASTIN Bulletin 21 (1991): 253-276.

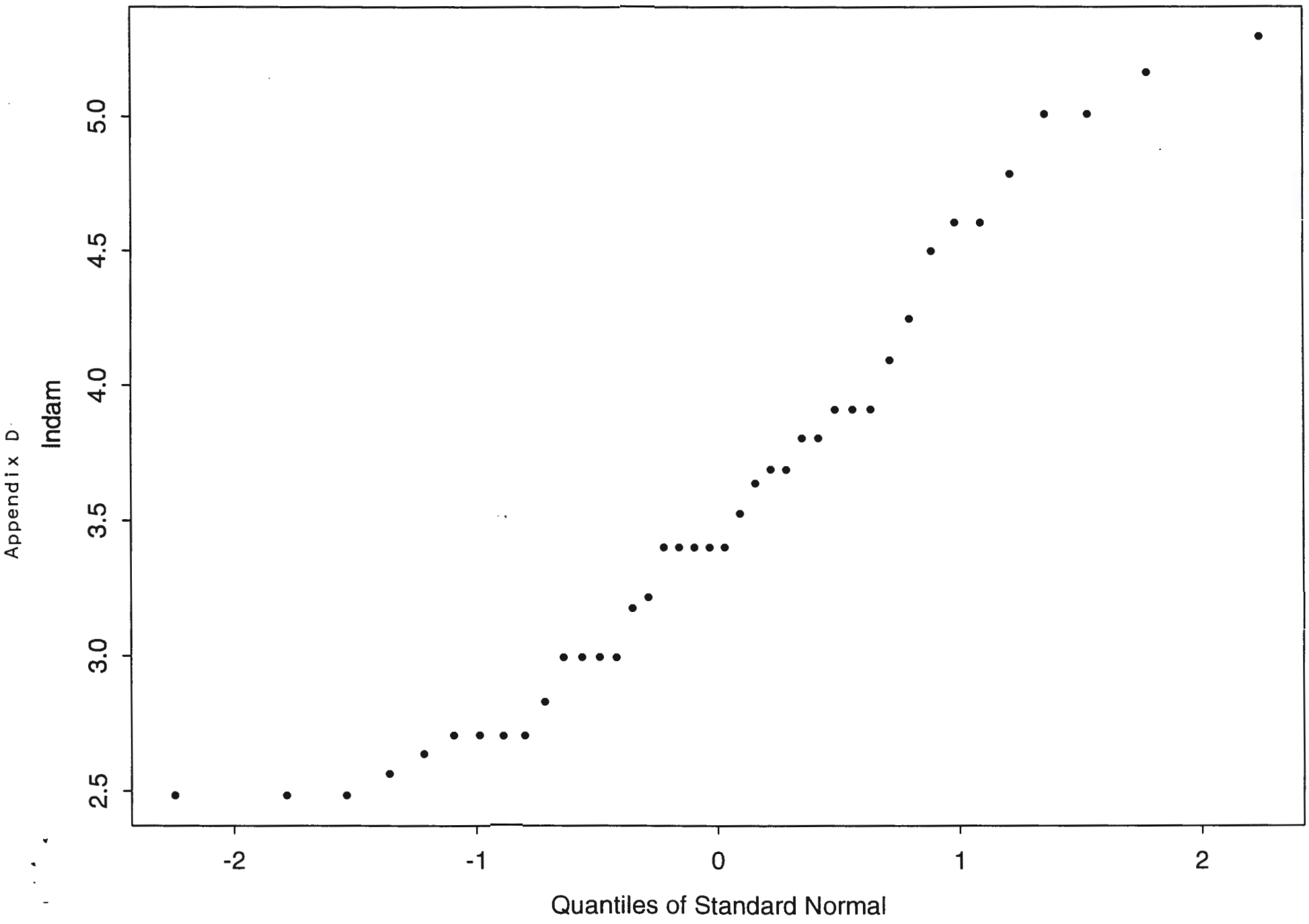
Rencher, Alvin C. Methods of Multivariate Analysis. New York:Wiley, 1995.


Histogram of Damages
Appendix A
0
2
4
6
8
10
12
14
0
50
100
150
200
forty[, 3]
Damages


Normal Probability Plot of Damages
Damages
fortyf[, 3]
200
150
100
50
-2
-1
0
1
2
Quantiles of Standard Normal

Appendix B


Histogram of Natural Log of Damages
Appendix C
0
2
4
6
8
10
2
3
4
5
lndam


Normal Probability Plot fo Natural Log of Damages
lndam
2.5
3.0
3.5
4.0
4.5
5.0
-2
-1
0
1
2
Quantiles of Standard Normal

Star Plots of Forty Tornados


1


2


3


4


5


6

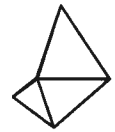

7


8


9

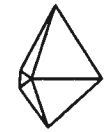

10

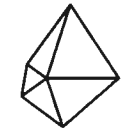

11


12


13


14


15

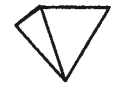

16


17


18


19


20


21


22


23


24


25


26


27


28


29


30


31

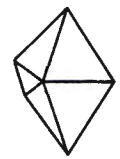

32


33


34

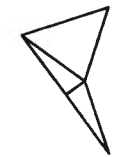

35


36


37

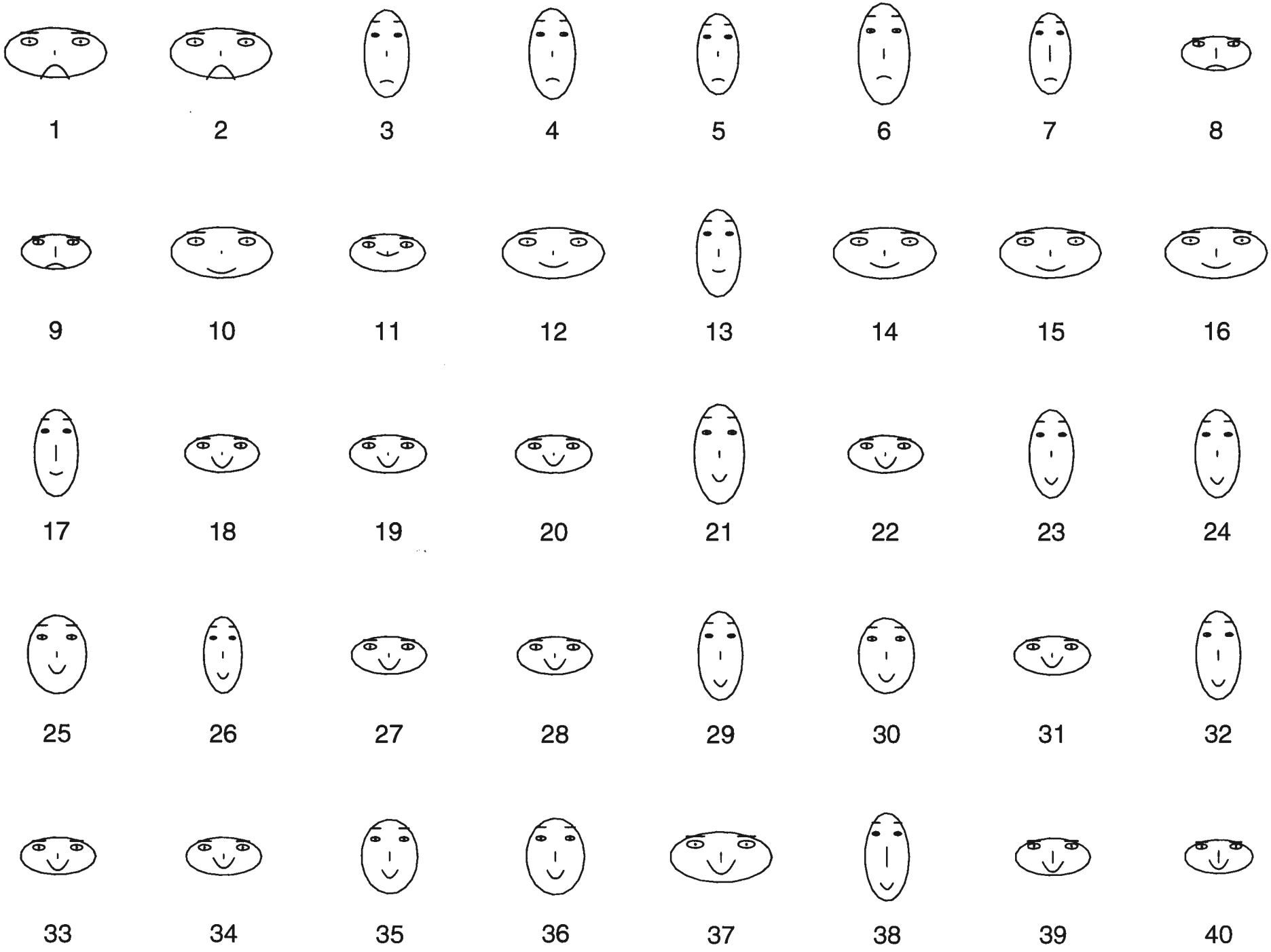

38


39


40

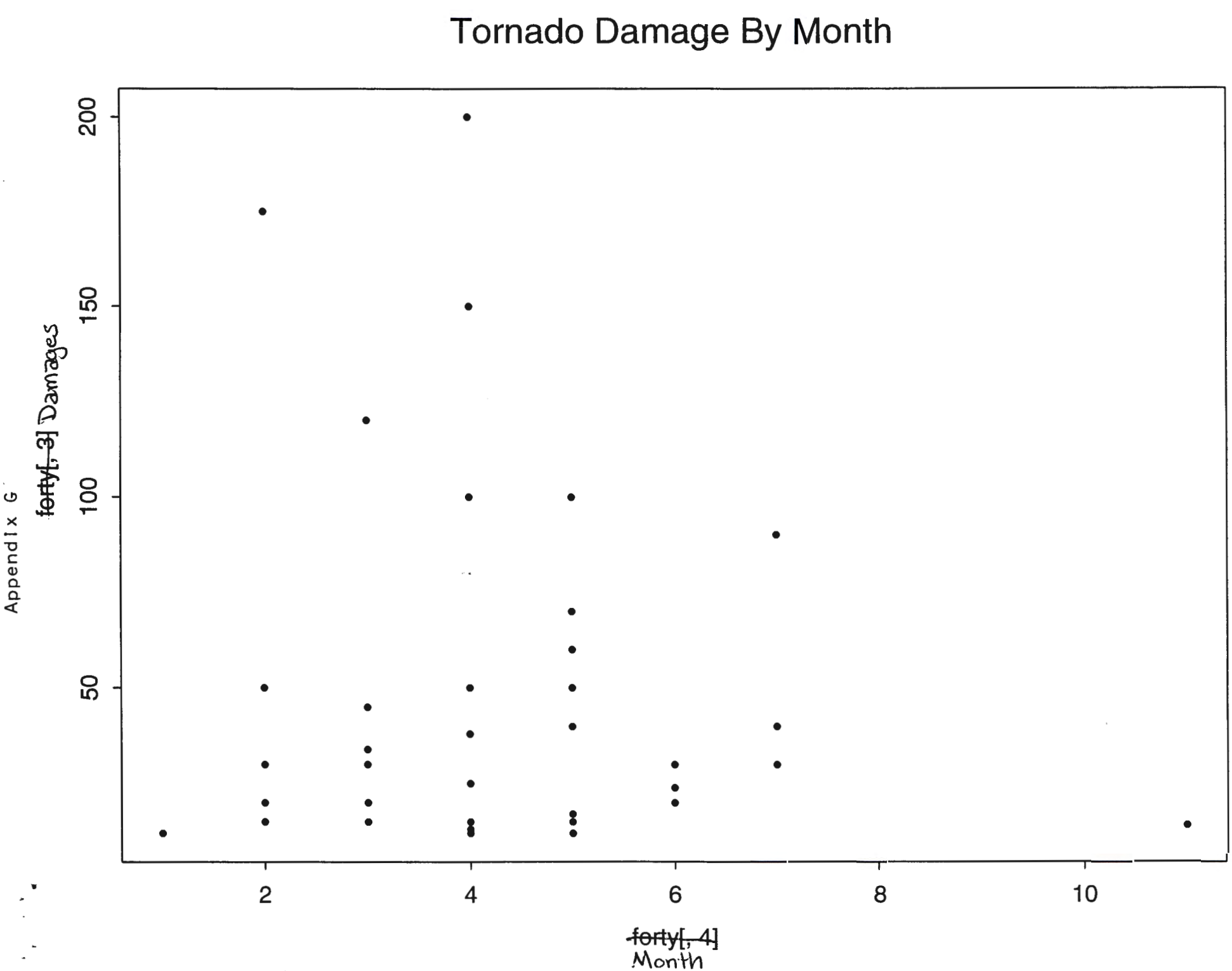
Appendix E

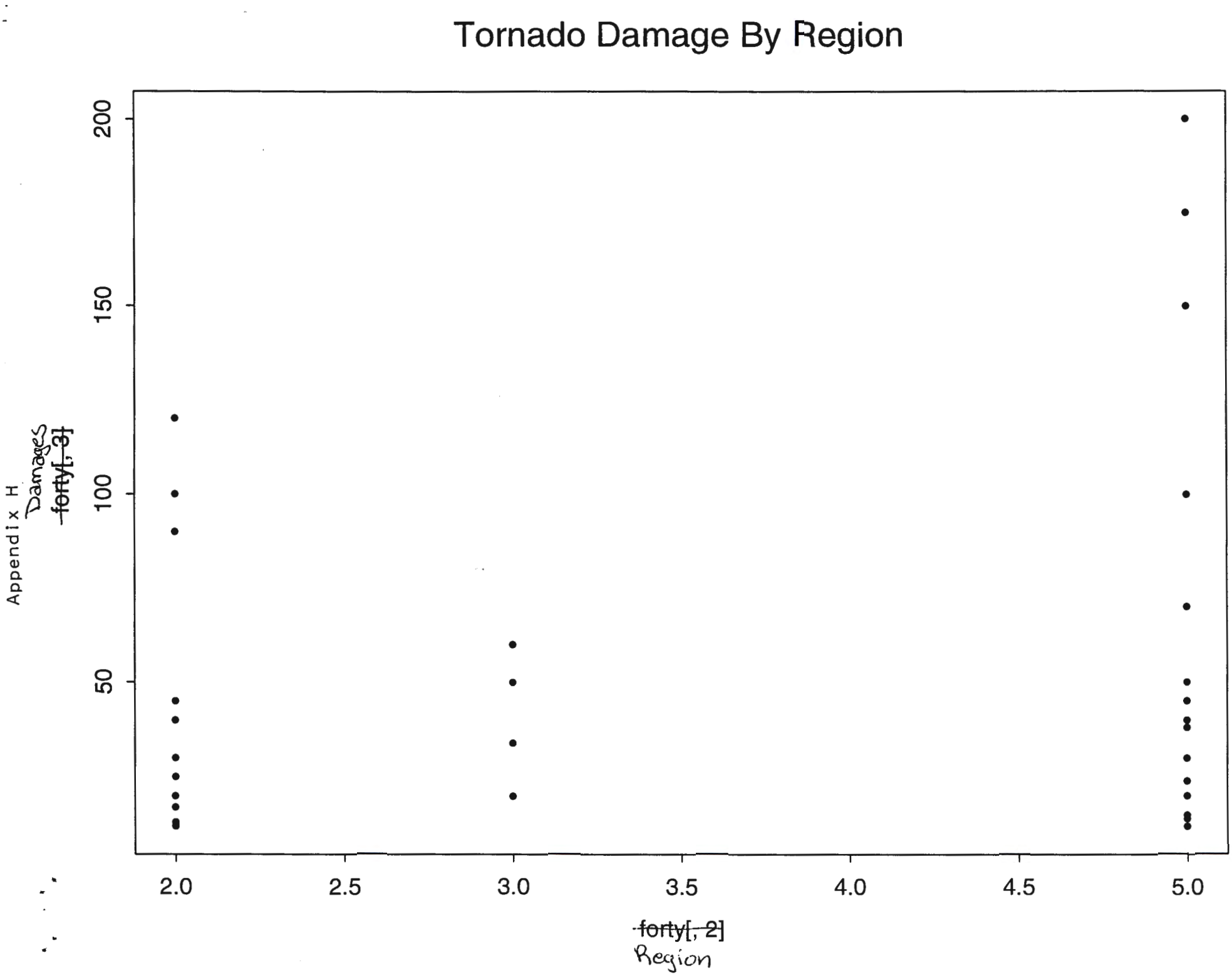
Faces of Forty Tornados



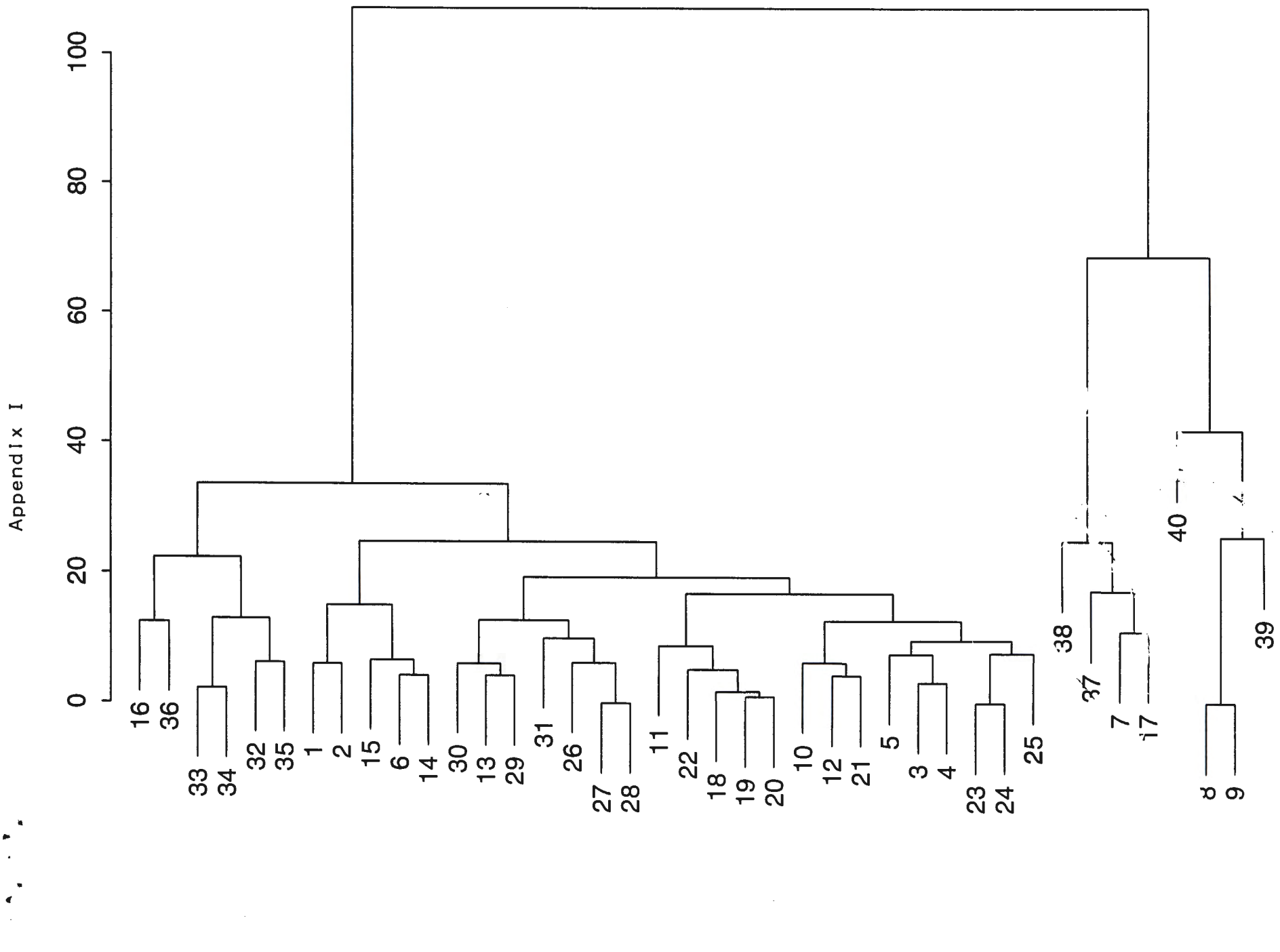
Appendix F

Appendix G


Tornado Damage By Month
forty[, 3] Damages
50
100
150
200
2
4
6
8
10
forty[, 4]
Month

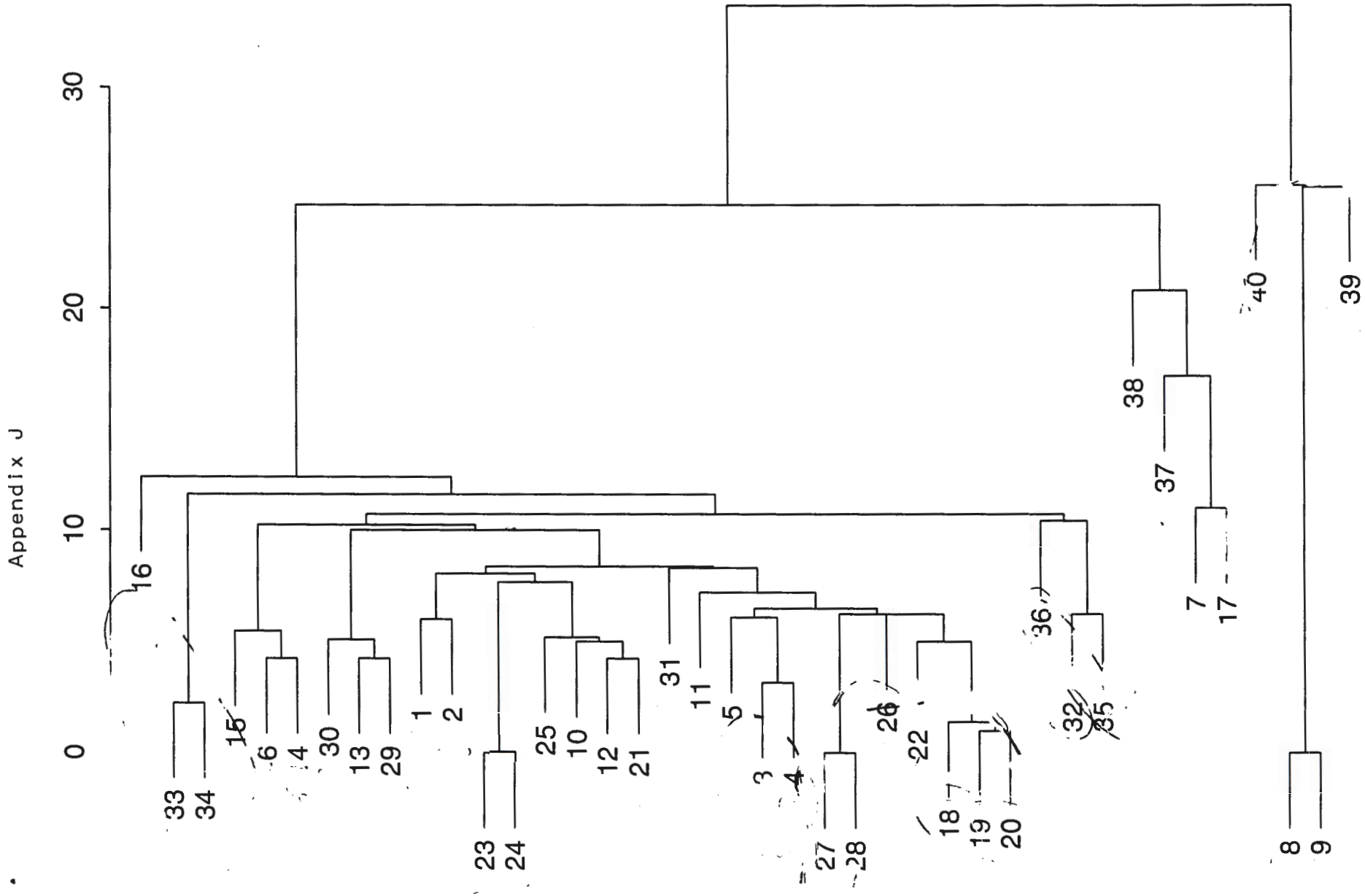

Tornado Damage By Region
Appendix H
Damages
forty[, 3]
200
150
100
50
2.0
2.5
3.0
3.5
4.0
4.5
5.0
forty[, 2]
Region

Average Linkage Dendogram-Euclidean Distance-Forty Tornados



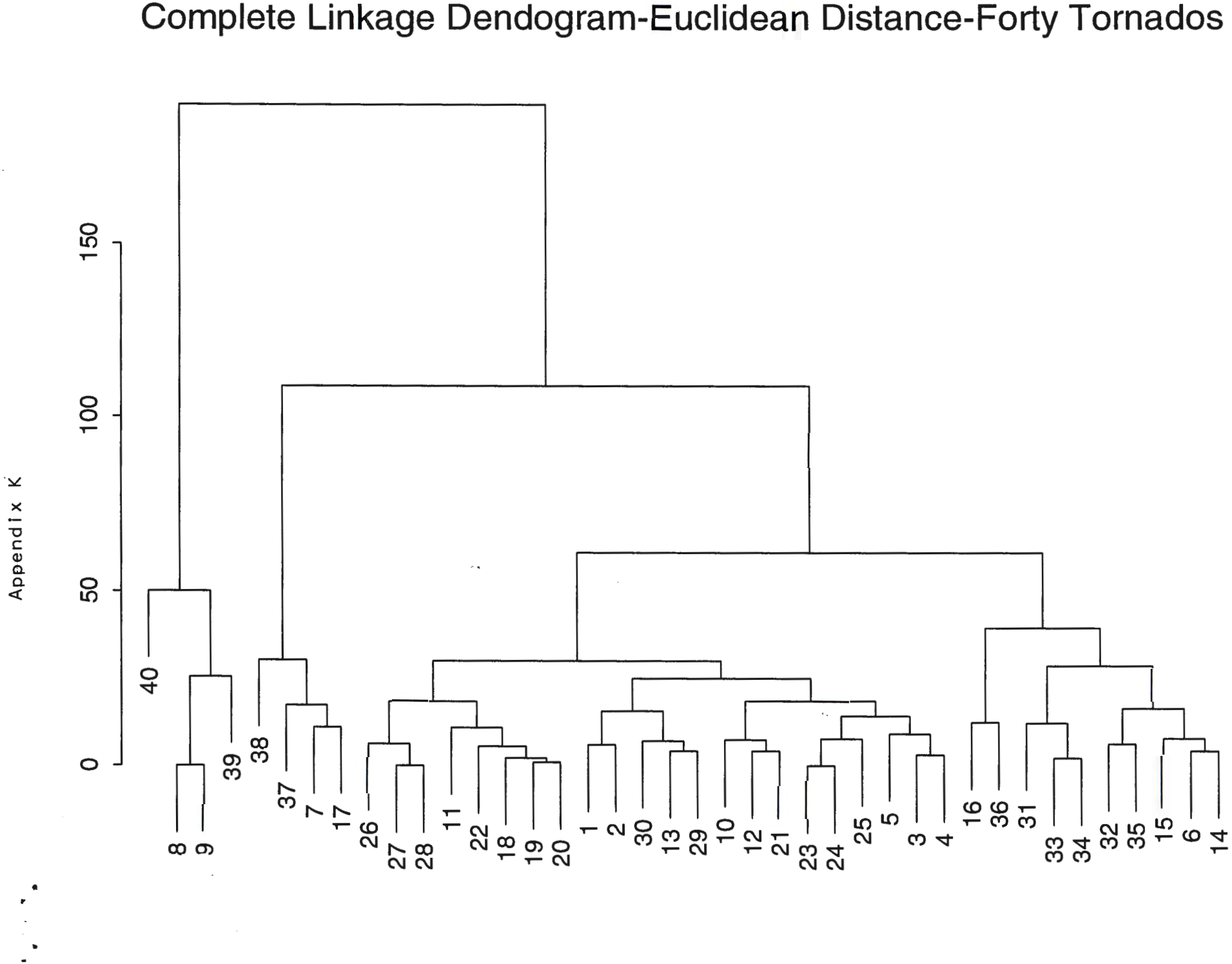
Appendix I

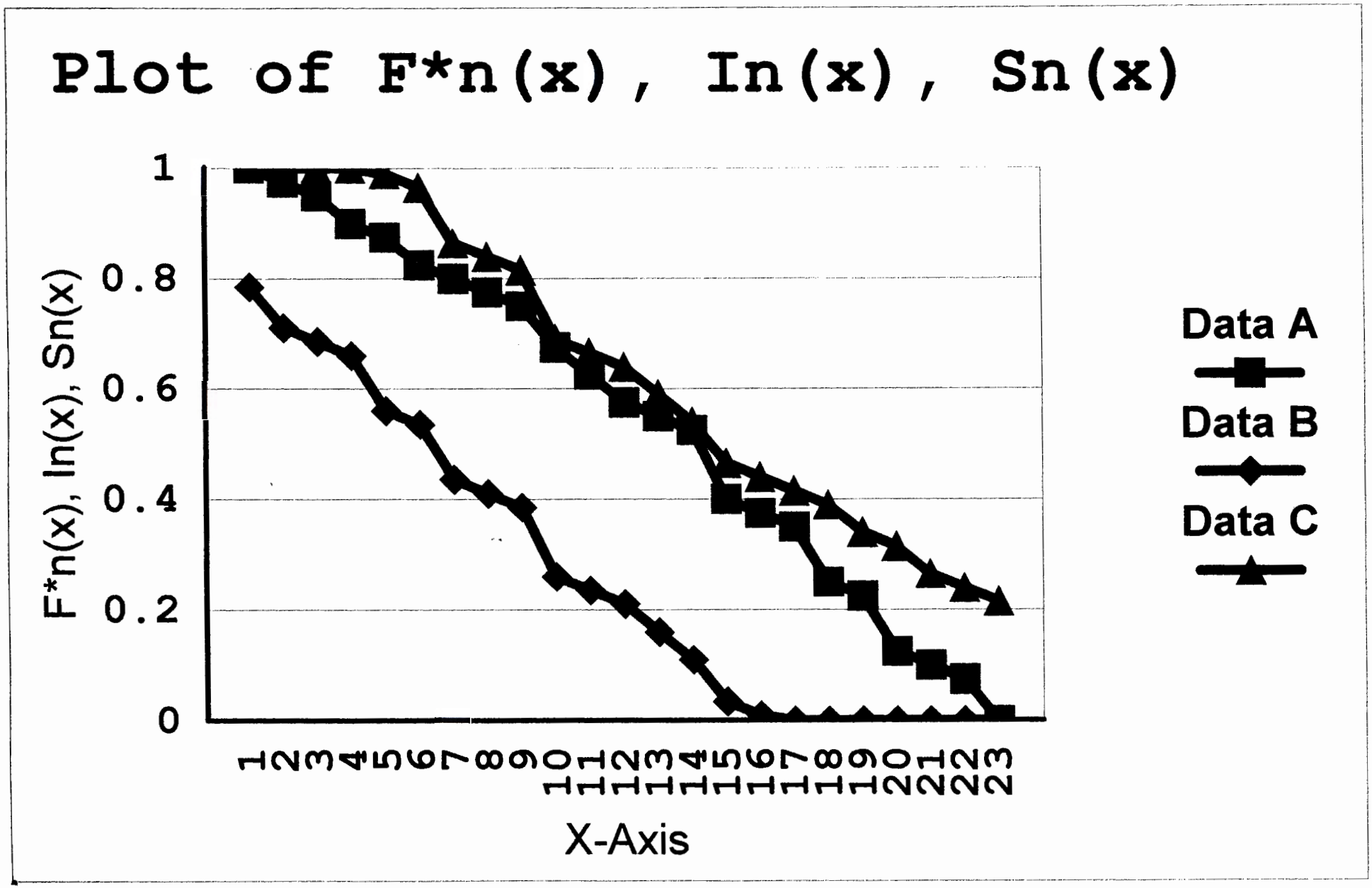
Single Linkage Dendogram-Euclidean Distance-Forty Tornados



Appendix J

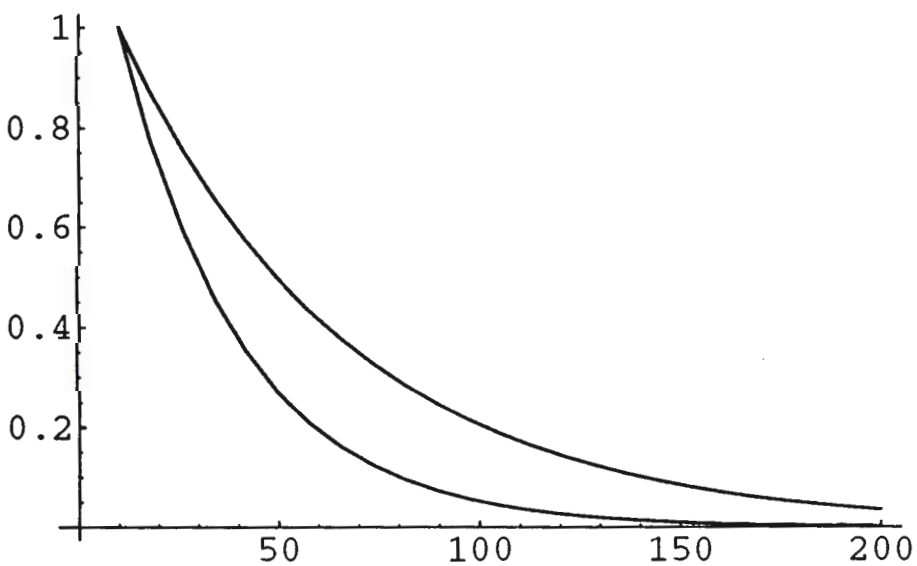
Appendix K


Complete Linkage Dendogram-Euclidean Distance-Forty Tornados
0
50
100
150
40
8
9
39
38
37
7
17
26
27
28
11
22
18
19
20
1
2
30
13
29
10
12
21
23
24
25
5
3
4
16
36
31
33
34
32
35
15
6
14


Plot of F*n(x), In(x), Sn(x)
F*n(x), In(x), Sn(x)
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
X-Axis
Data A
Data B
Data C

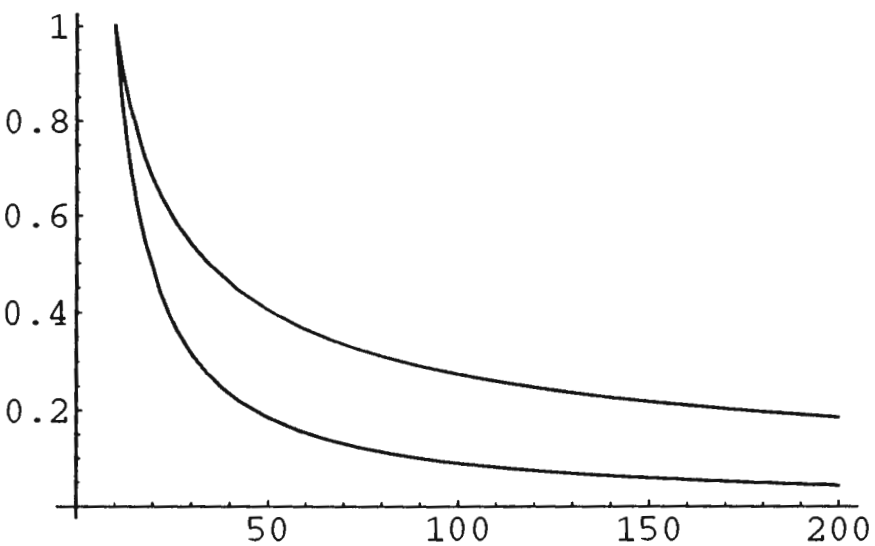
Exponential

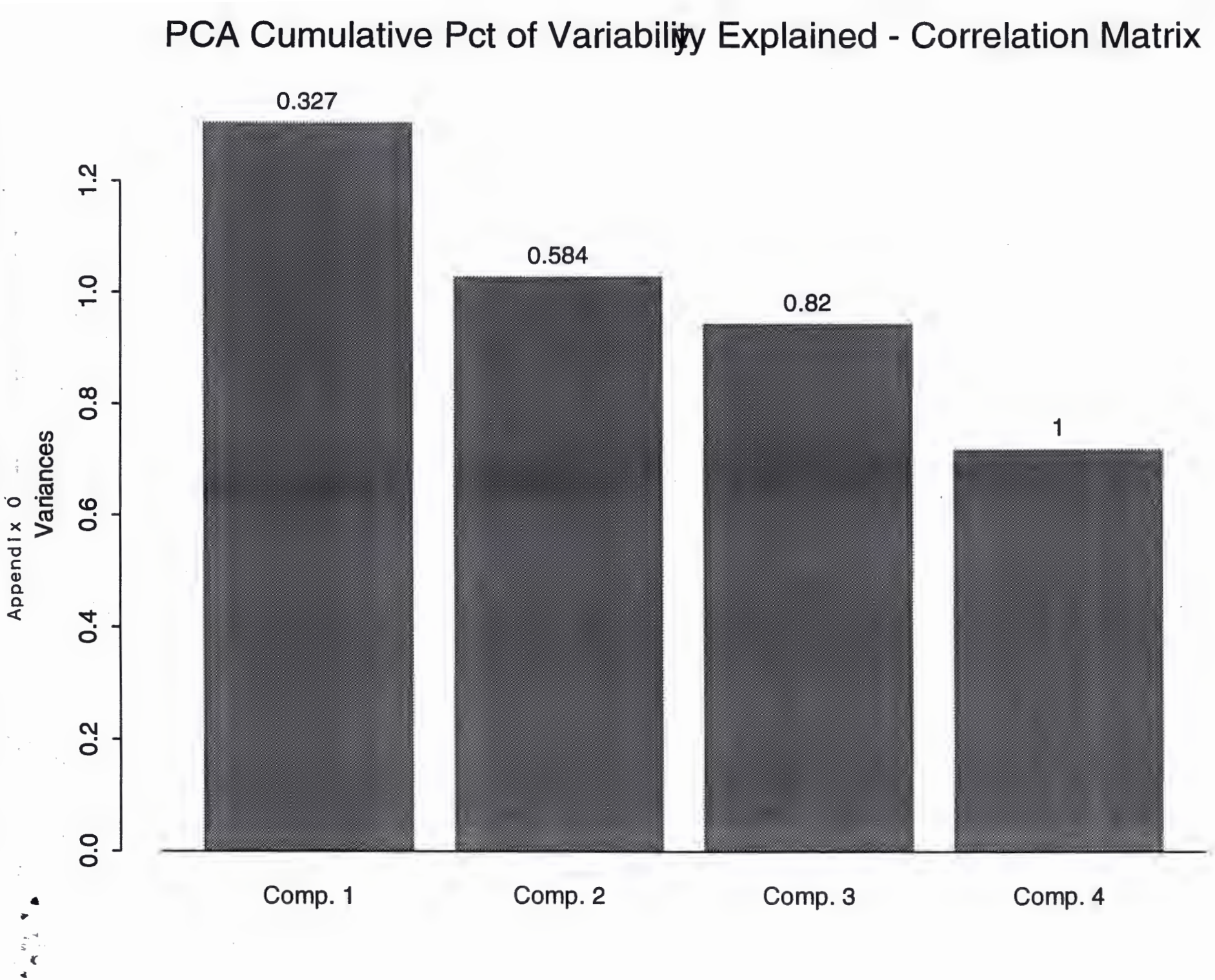
```
In[8]:= Plot[{Exp[-.03278 * (x - 10)], Exp[-.01757 * (x - 10)]}, {x, 10, 200}]
```

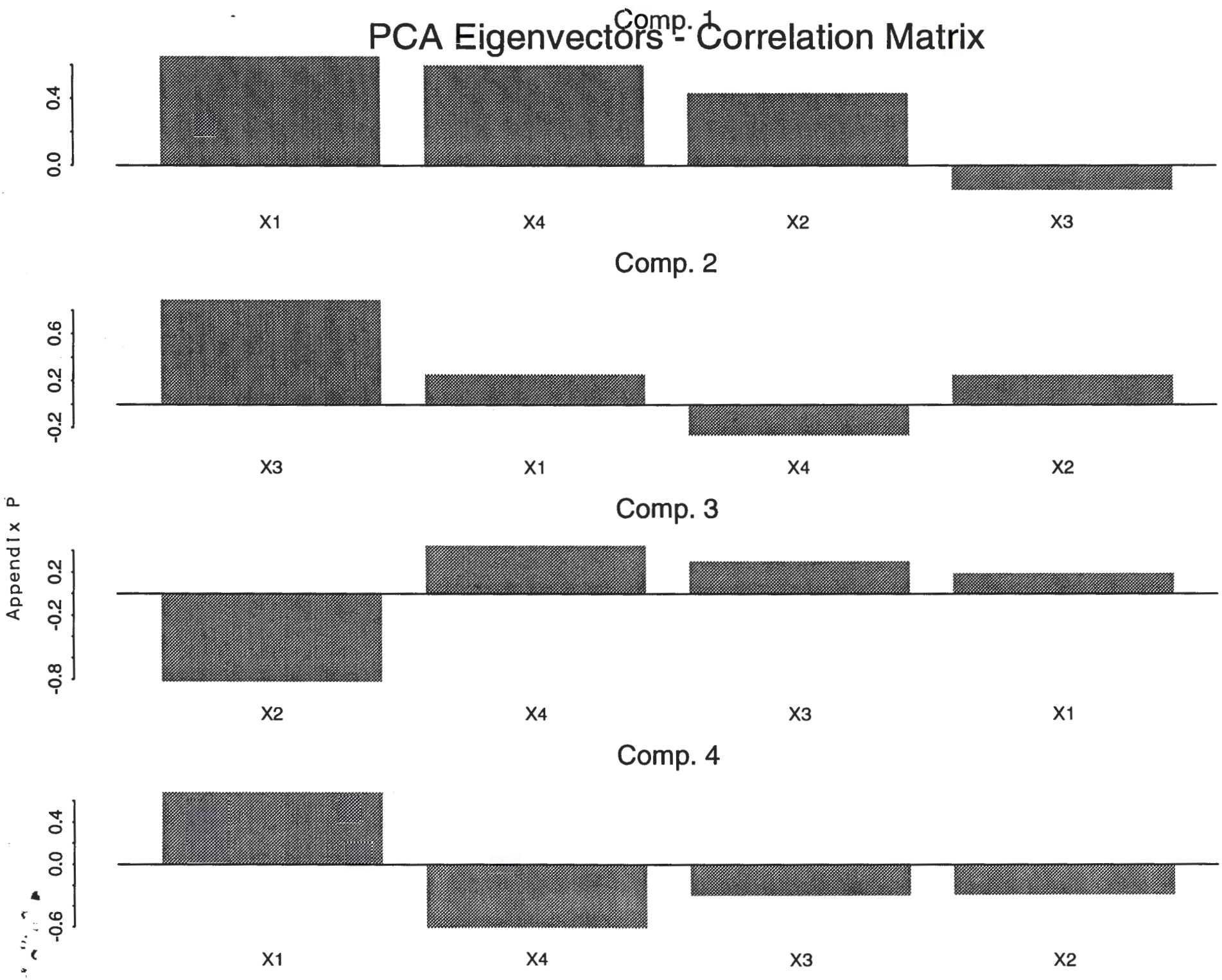


Pareto

In[11]:= **Plot[{(10 / x) ^ .56225, (10 / x) ^ 1.0487}, {x, 10, 200}]**

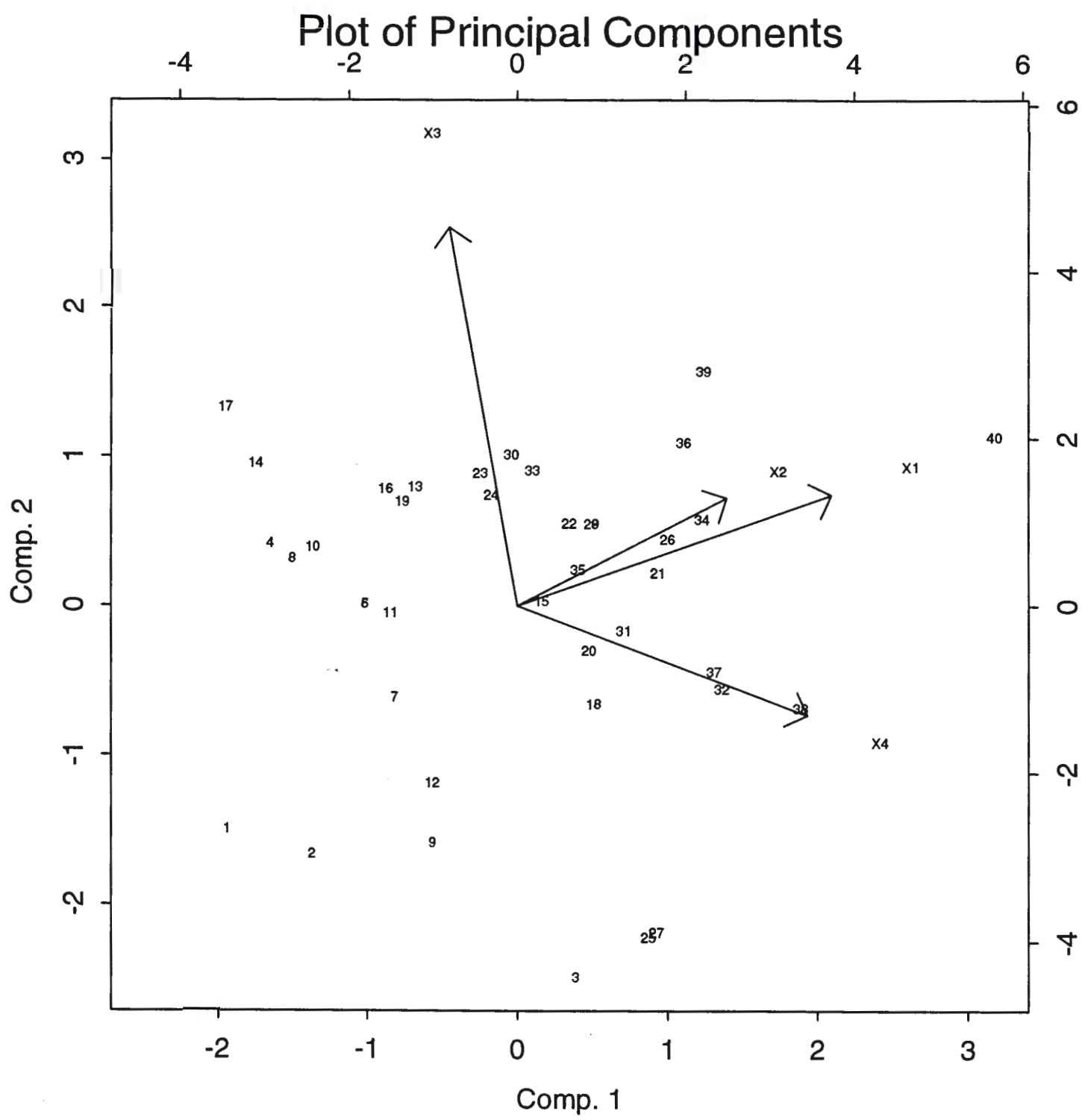


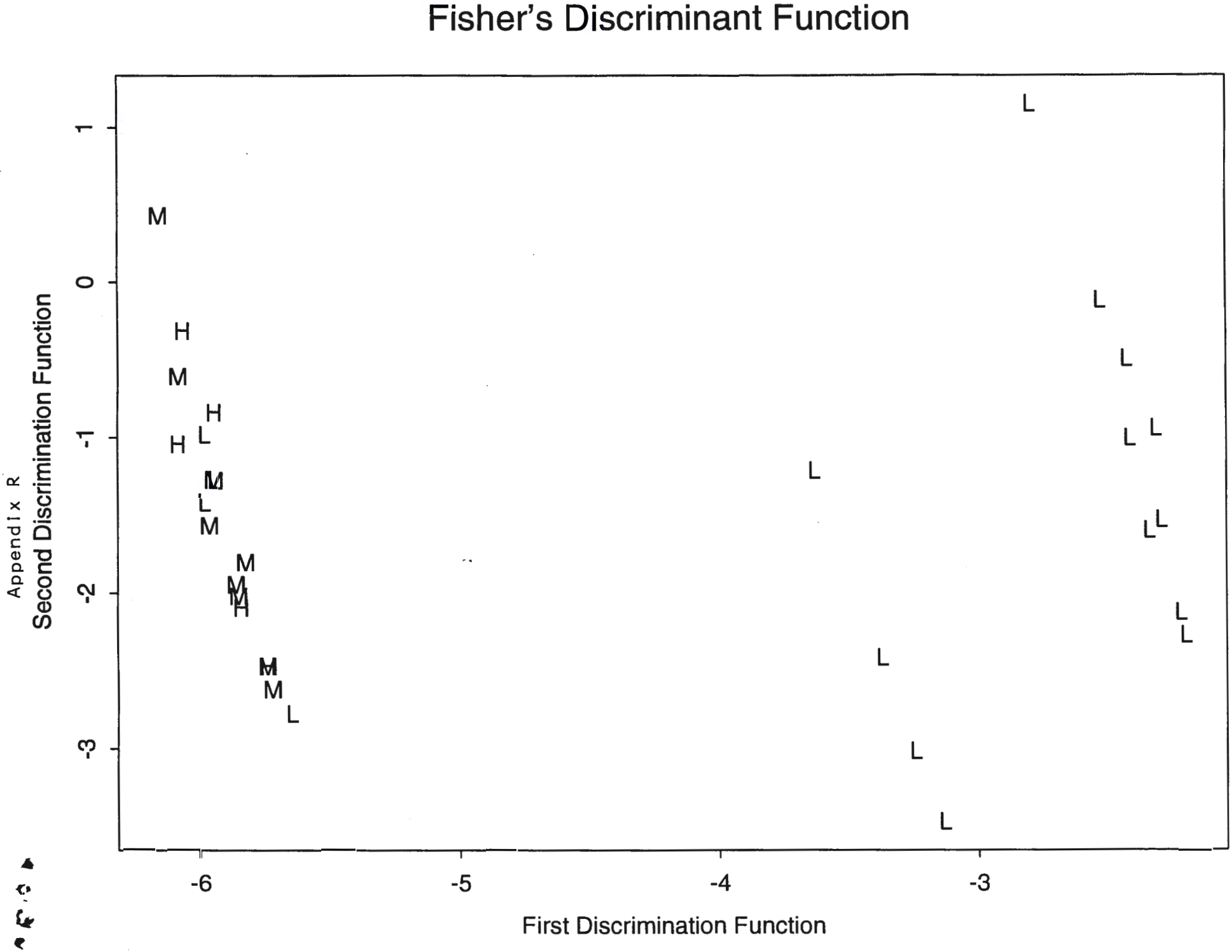

PCA Cumulative Pct of Variability Explained - Correlation Matrix
0.327
0.584
0.82
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Variances
Comp. 1
Comp. 2
Comp. 3
Comp. 4


Comp. 1
PCA Eigenvectors - Correlation Matrix
0.0
0.4
X1
X4
X2
X3
Comp. 2
-0.2
0.2
0.6
X3
X1
X4
X2
Comp. 3
-0.8
-0.2
0.2
X2
X4
X3
X1
Comp. 4
-0.6
0.0
0.4
X1
X4
X3
X2

Appendix P

Appendix Q


Plot of Principal Components
Comp. 1
Comp. 2
X1
X2
X3
X4
-4
-2
0
2
4
6
-2
-1
0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40


Fisher's Discriminant Function
Second Discrimination Function
First Discrimination Function
1
0
-1
-2
-3
-6
-5
-4
-3

Appendix R