Building statistical models by visualization

Tom Minka
CMU Statistics Dept

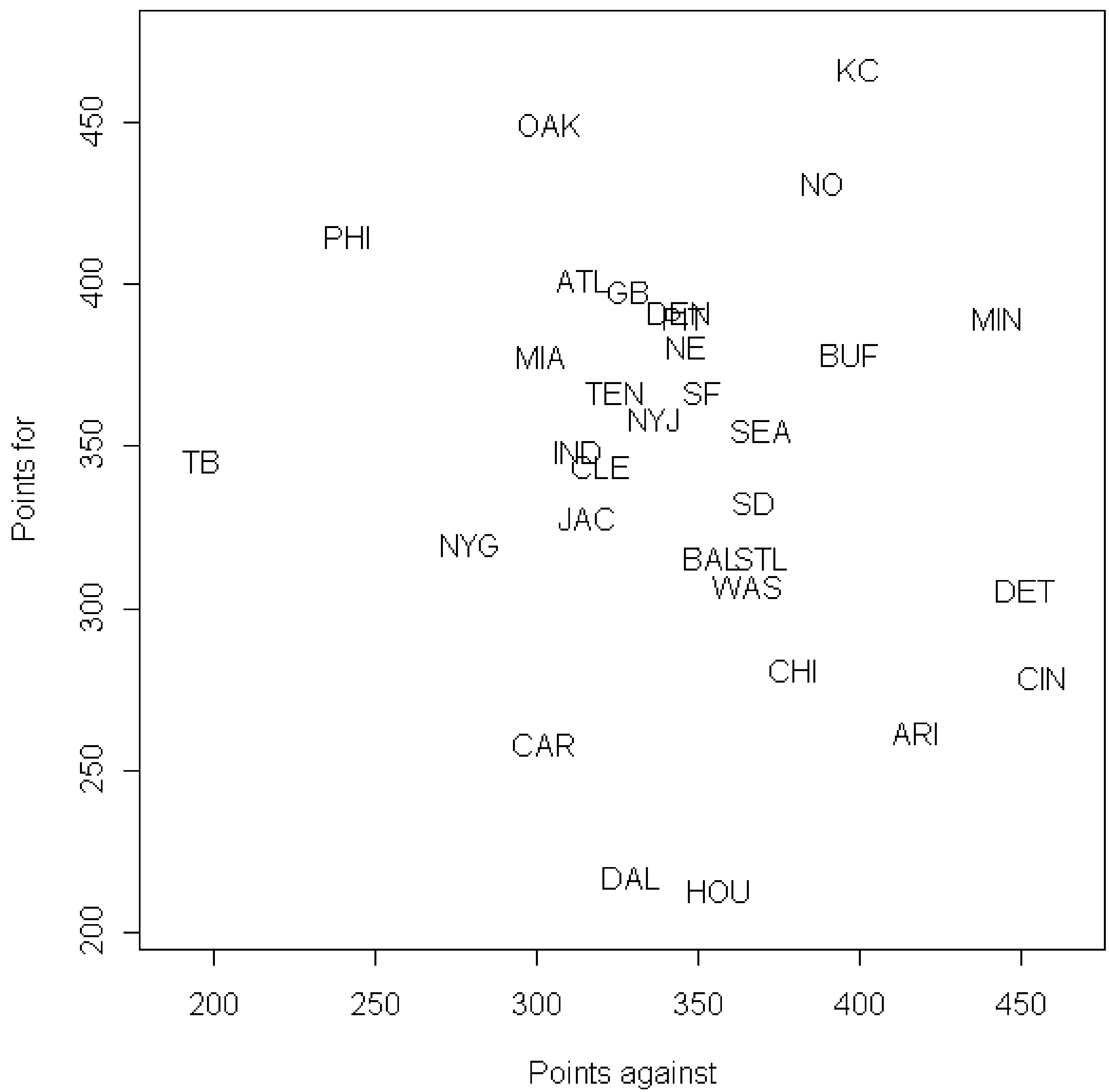
Outline

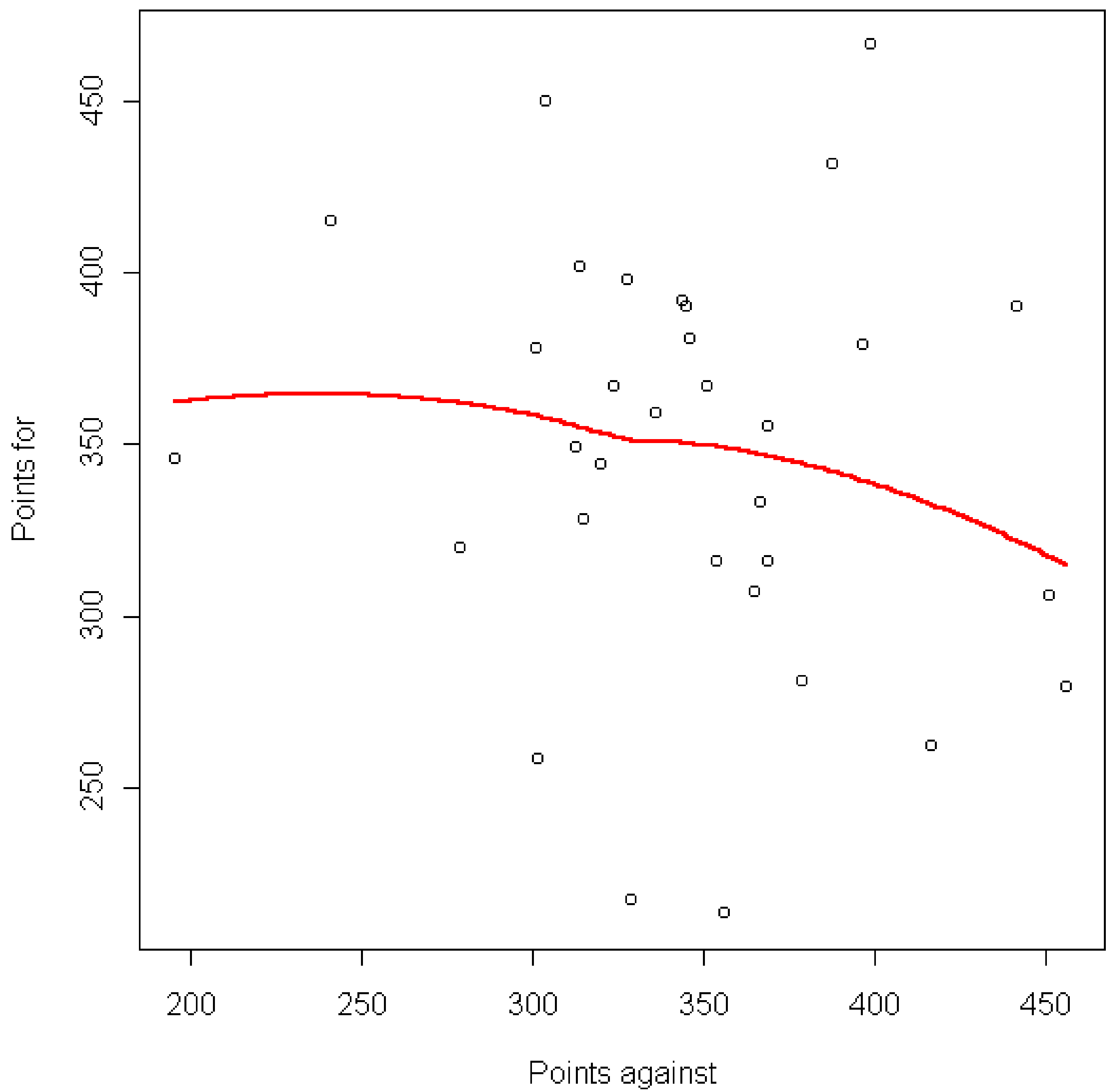
- Scatterplots
  - independence, causality
- QQ plots
  - distribution checking
- Residual plots
  - linearity, outliers
- Projections for regression
  - additivity
- Projections for classification
  - linearity

Football statistics

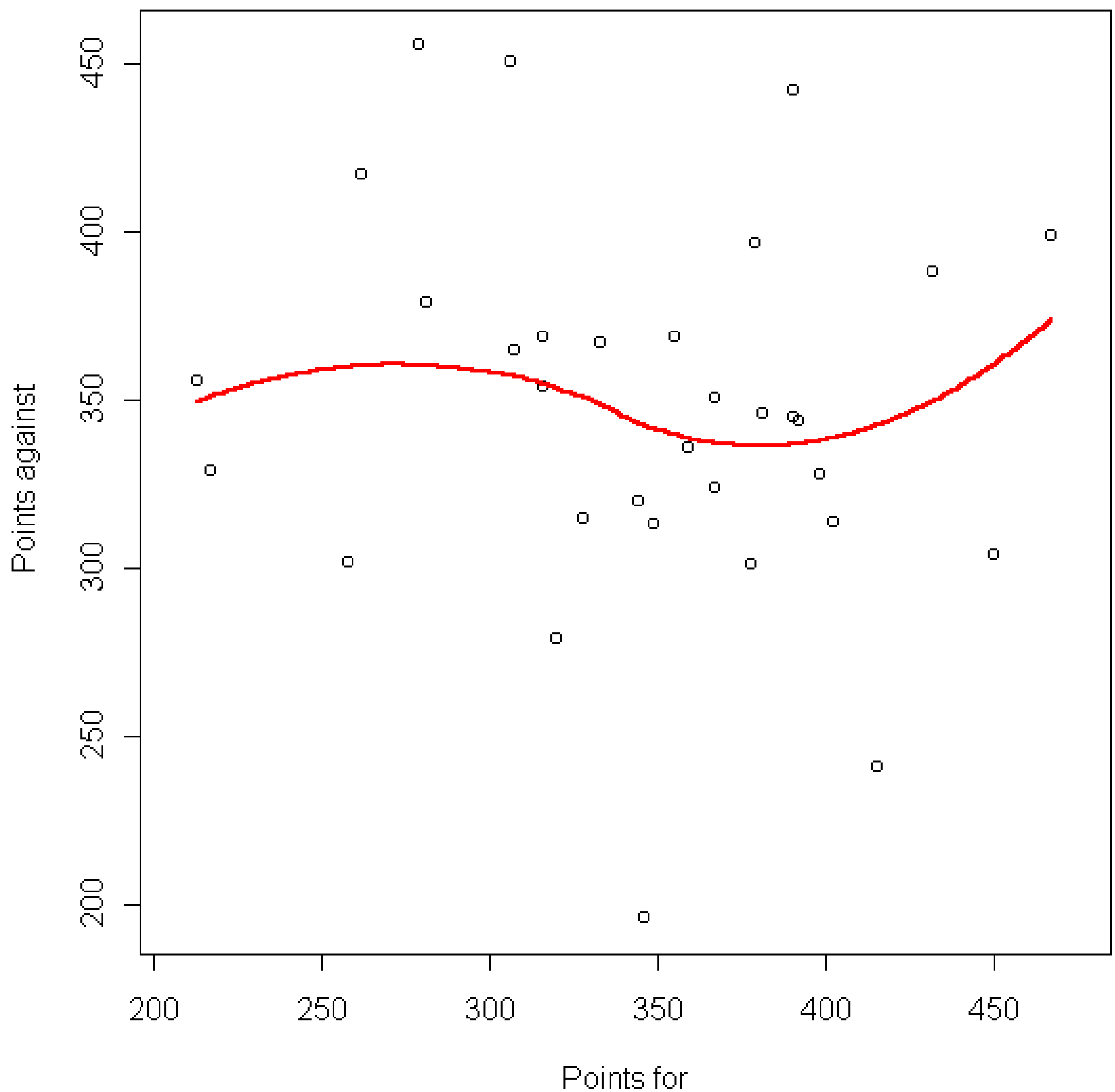
Is there a better representation?

	W	L	T	PF	PA
CIN	2	14	0	279	456
OAK	11	5	0	450	304
GB	12	4	0	398	328
SEA	7	9	0	355	369
JAC	6	10	0	328	315
NO	9	7	0	432	388
KC	8	8	0	467	399
TB	12	4	0	346	196
MIN	6	10	0	390	442
TEN	11	5	0	367	324
CAR	7	9	0	258	302
NYJ	9	7	0	359	336
NE	9	7	0	381	346
STL	7	9	0	316	369
…					


KC
OAK
NO
PHI
ATL
GB
DEN
PIT
MIN
NE
MIA
BUF
TEN
SF
NYJ
SEA
TB
IND
CLE
SD
JAC
NYG
BAL
STL
WAS
DET
CHI
CIN
CAR
ARI
DAL
HOU
Points for
Points against
200
250
300
350
400
450



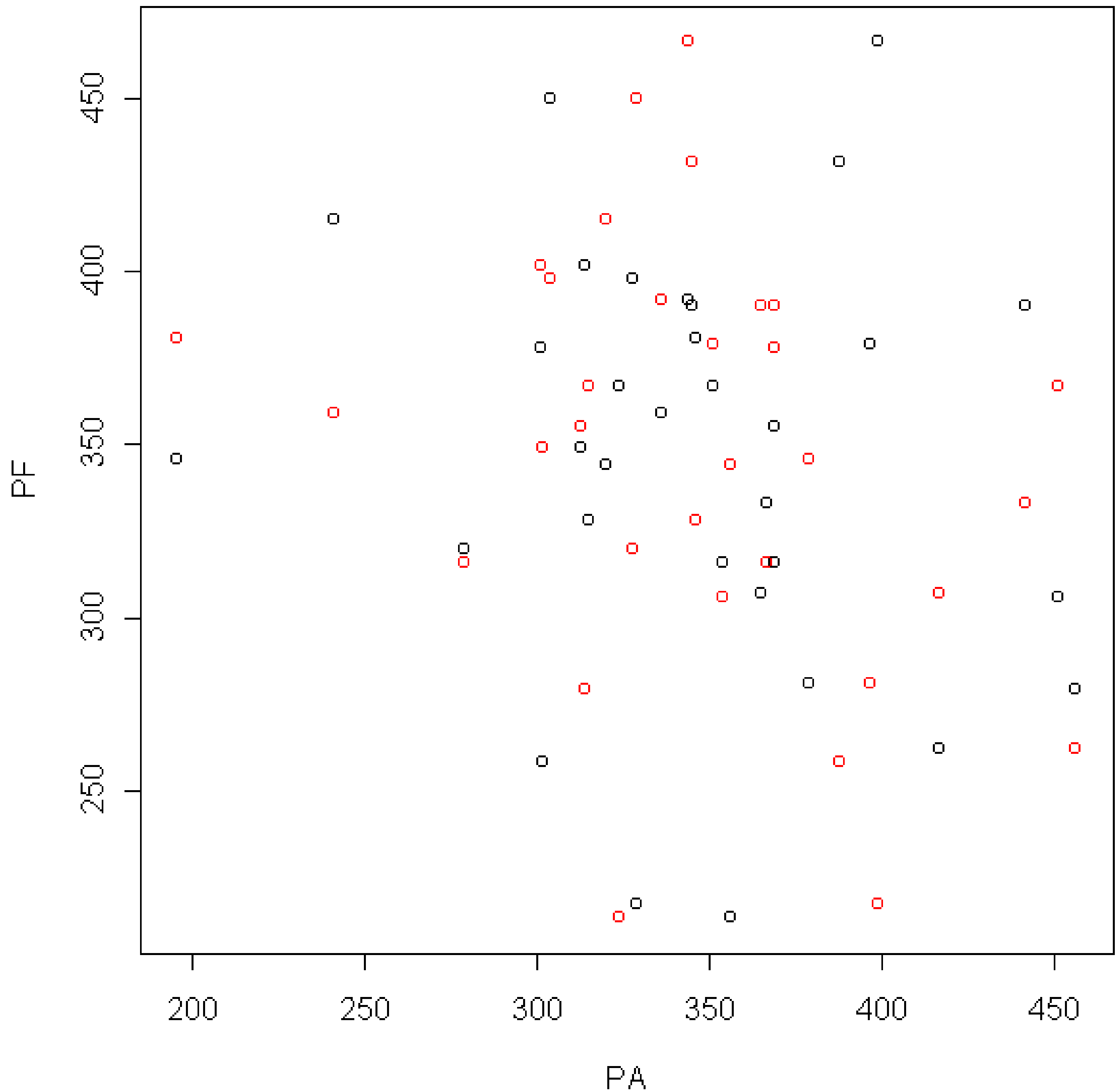
Points for
Points against
200
250
300
350
400
450
250
300
350
400
450



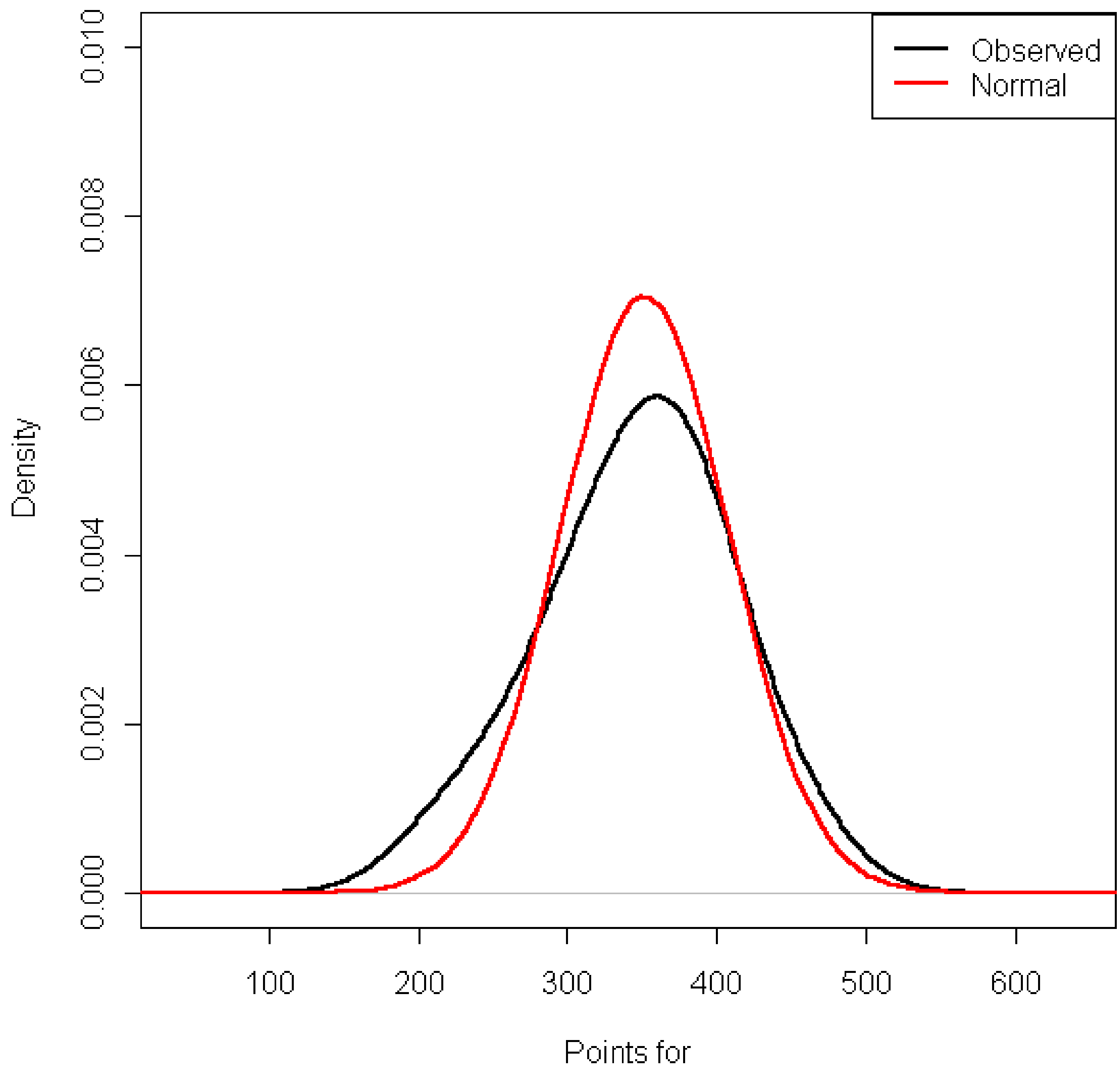
450
400
350
300
250
200
Points against
200
250
300
350
400
450
Points for

Visual independence test

- “Permutation test”
- Randomly pair x values with y values
- If the distribution looks different from the original, the variables are dependent
- No distributional assumptions required

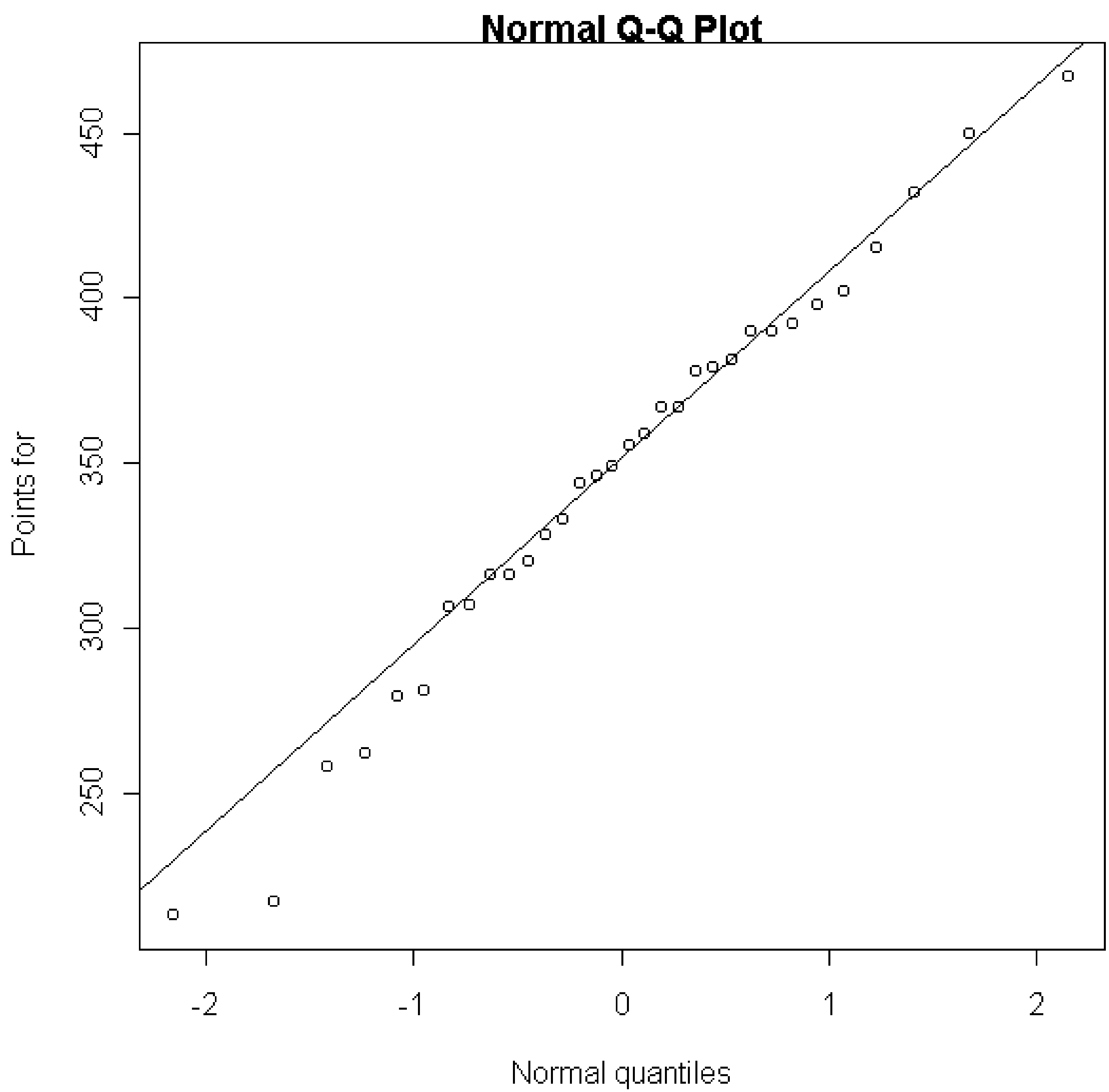


PF
450
400
350
300
250
200
250
300
350
400
450
PA


Observed
Normal
Density
0.000
0.002
0.004
0.006
0.008
0.010
100
200
300
400
500
600
Points for

Comparing distributions

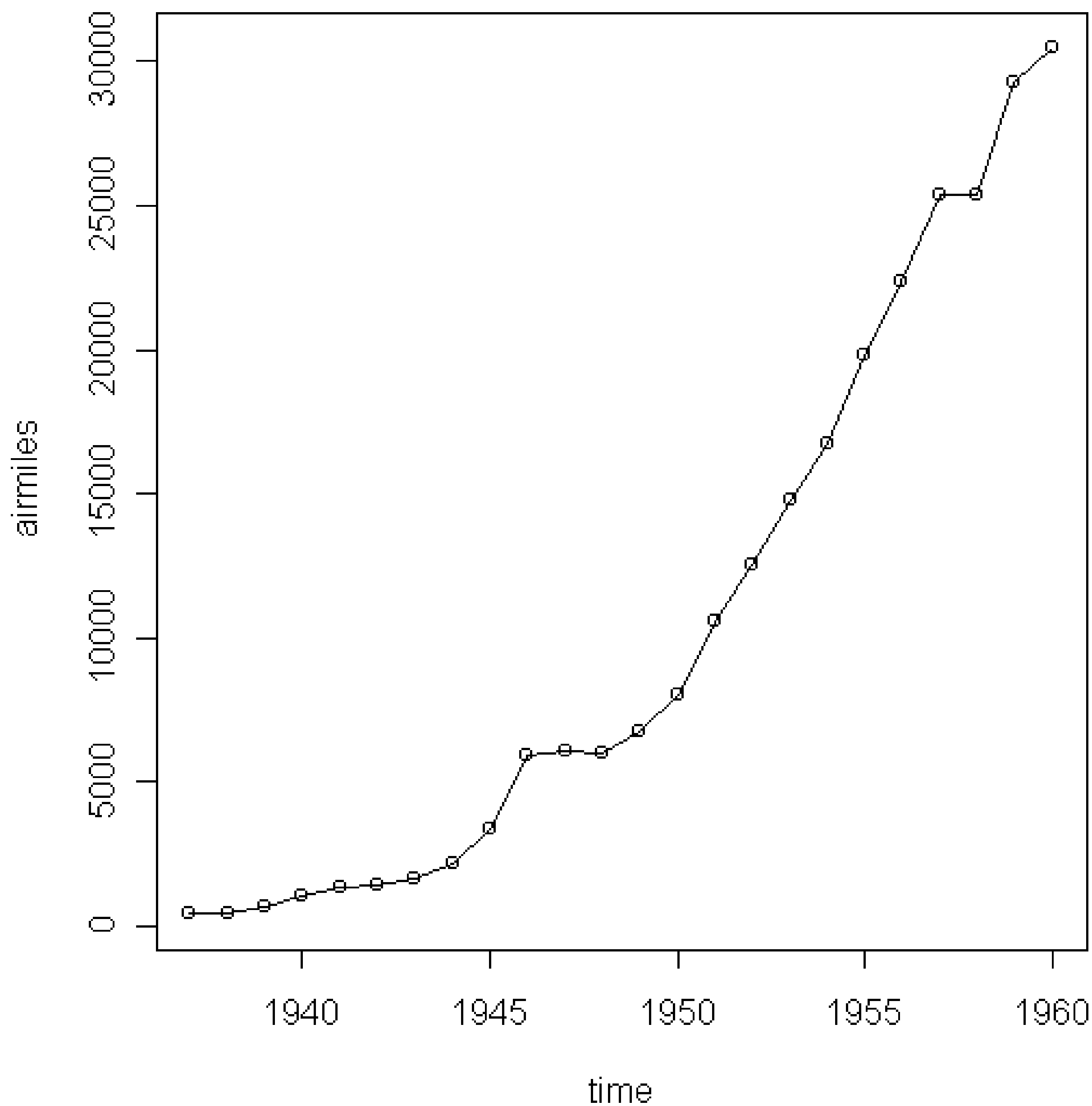
- “Quantile-quantile plot”
- A pseudo-scatterplot for unpaired data
- Quantile of x = fraction of points < x
- Plot quantile q in one set against quantile q in the other set, for all q
- Tells you how to transform one variable to have the distribution of the other

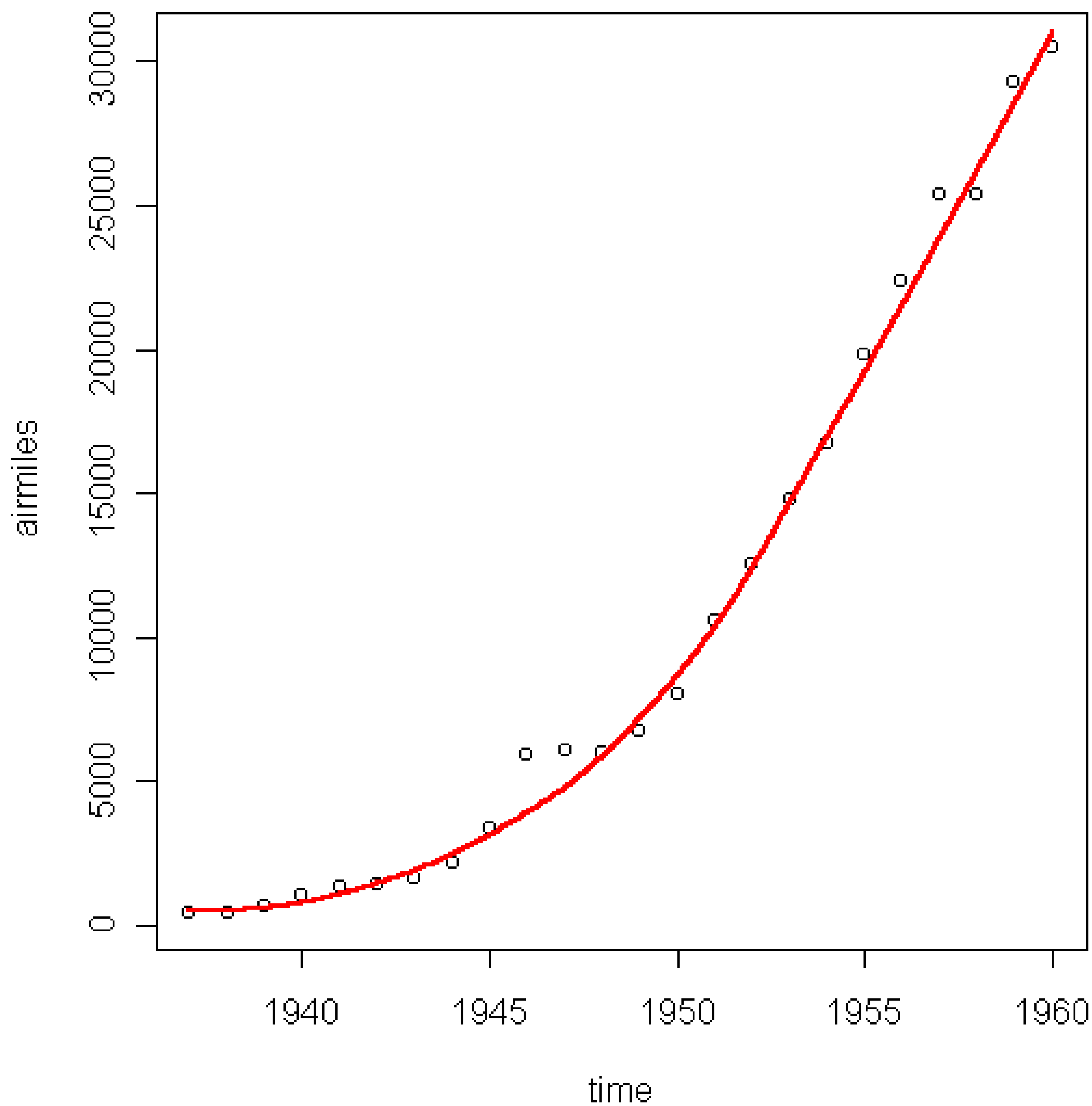

Normal Q-Q Plot
Points for
250
300
350
400
450
-2
-1
0
1
2
Normal quantiles

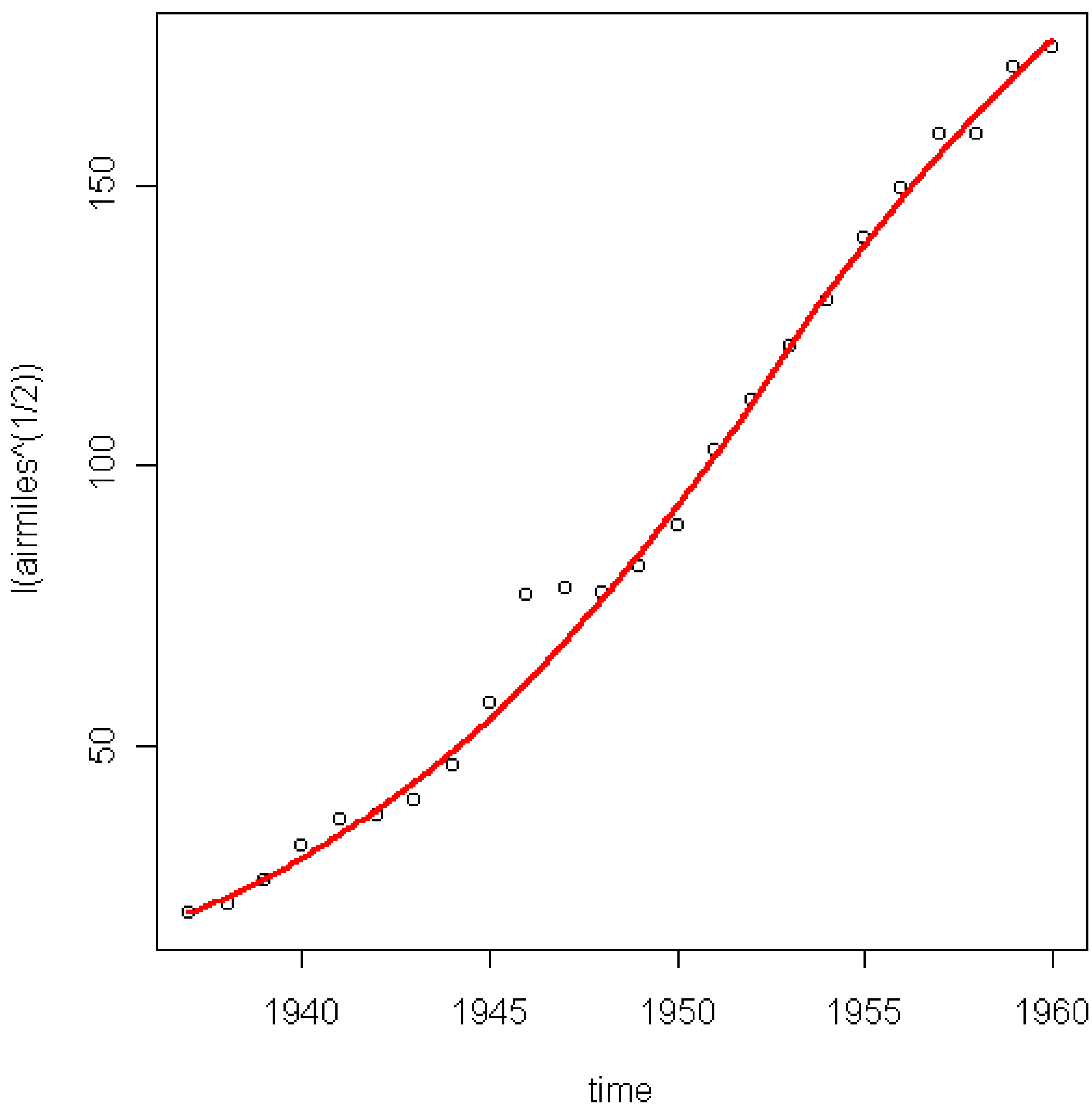
Regression models

How to do regression visually:

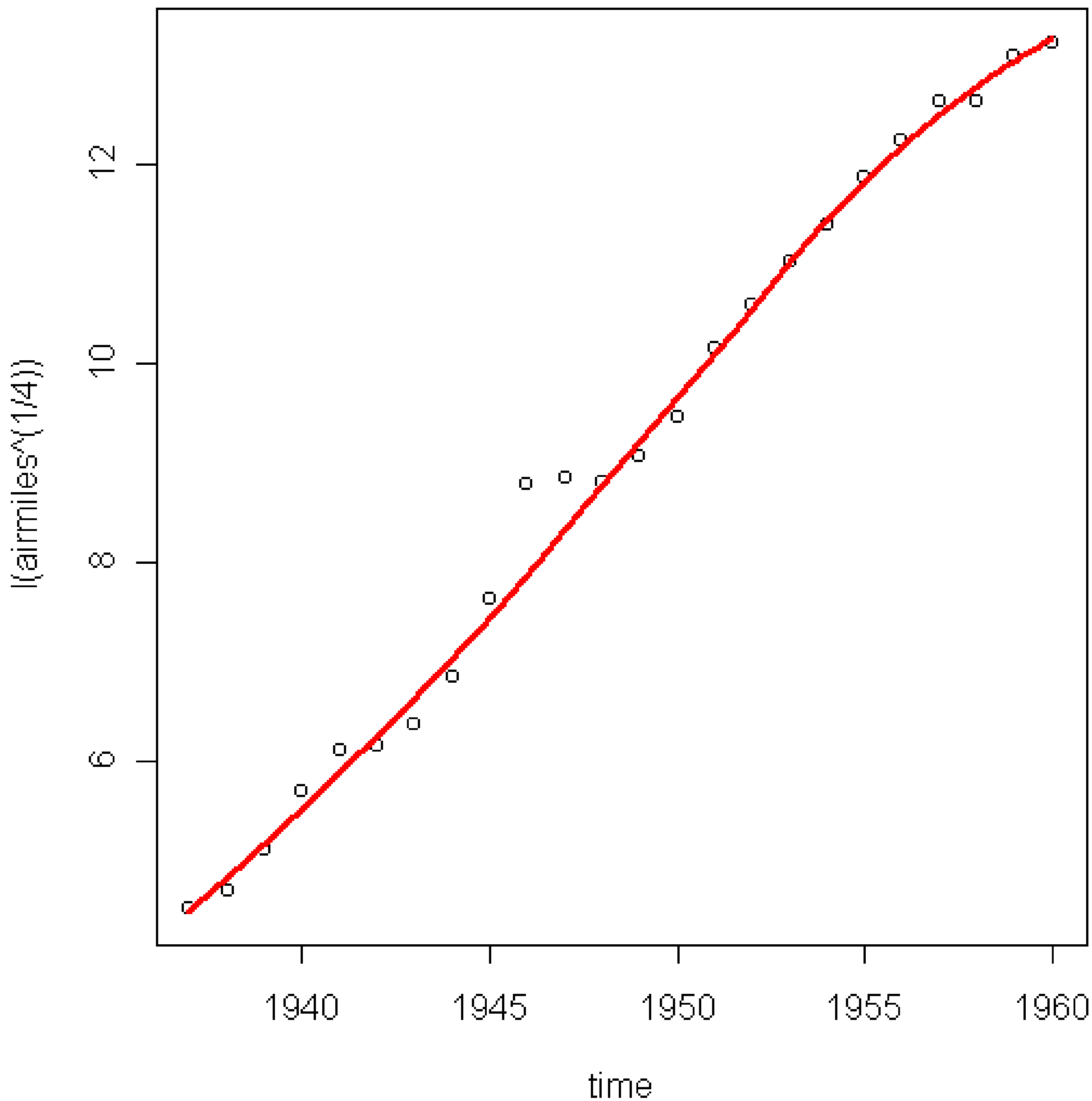
- Transform to make the picture simpler
- Fit a simple model
- Use residuals to suggest more complex models, outliers to remove
- Iterate

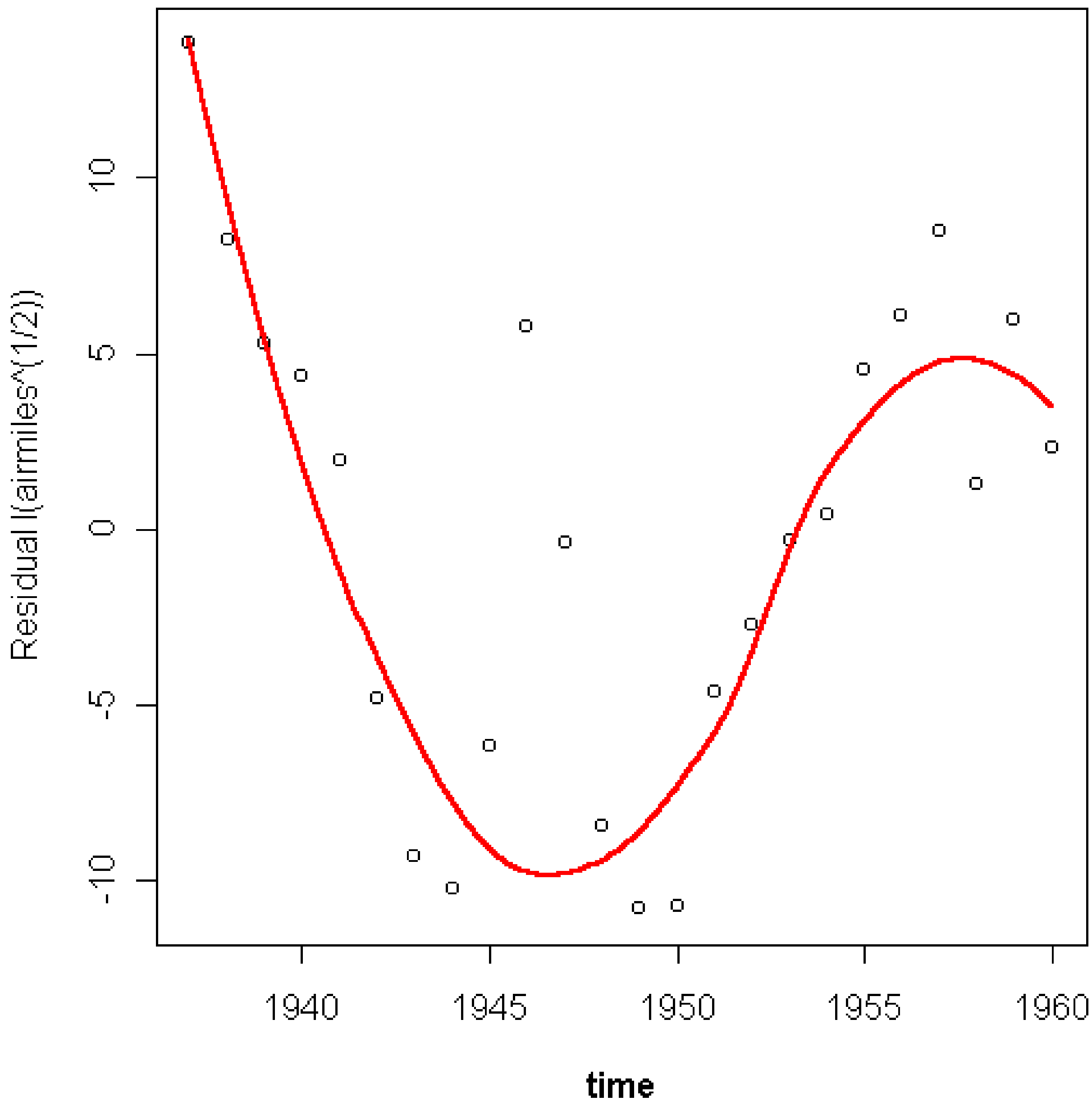

airmiles
0
5000
10000
15000
20000
25000
30000
1940
1945
1950
1955
1960
time


airmiles
time
0
5000
10000
15000
20000
25000
30000
1940
1945
1950
1955
1960

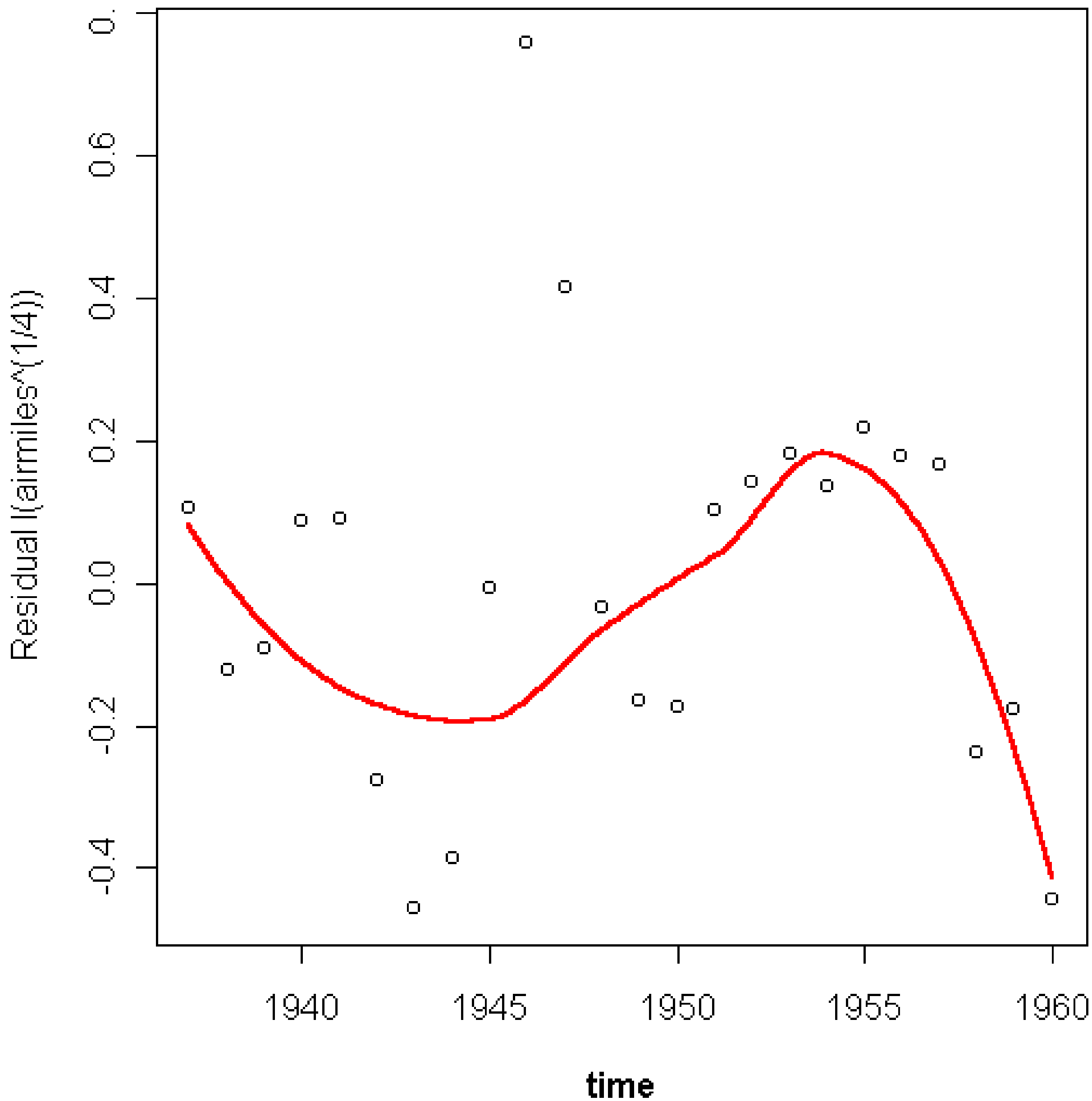


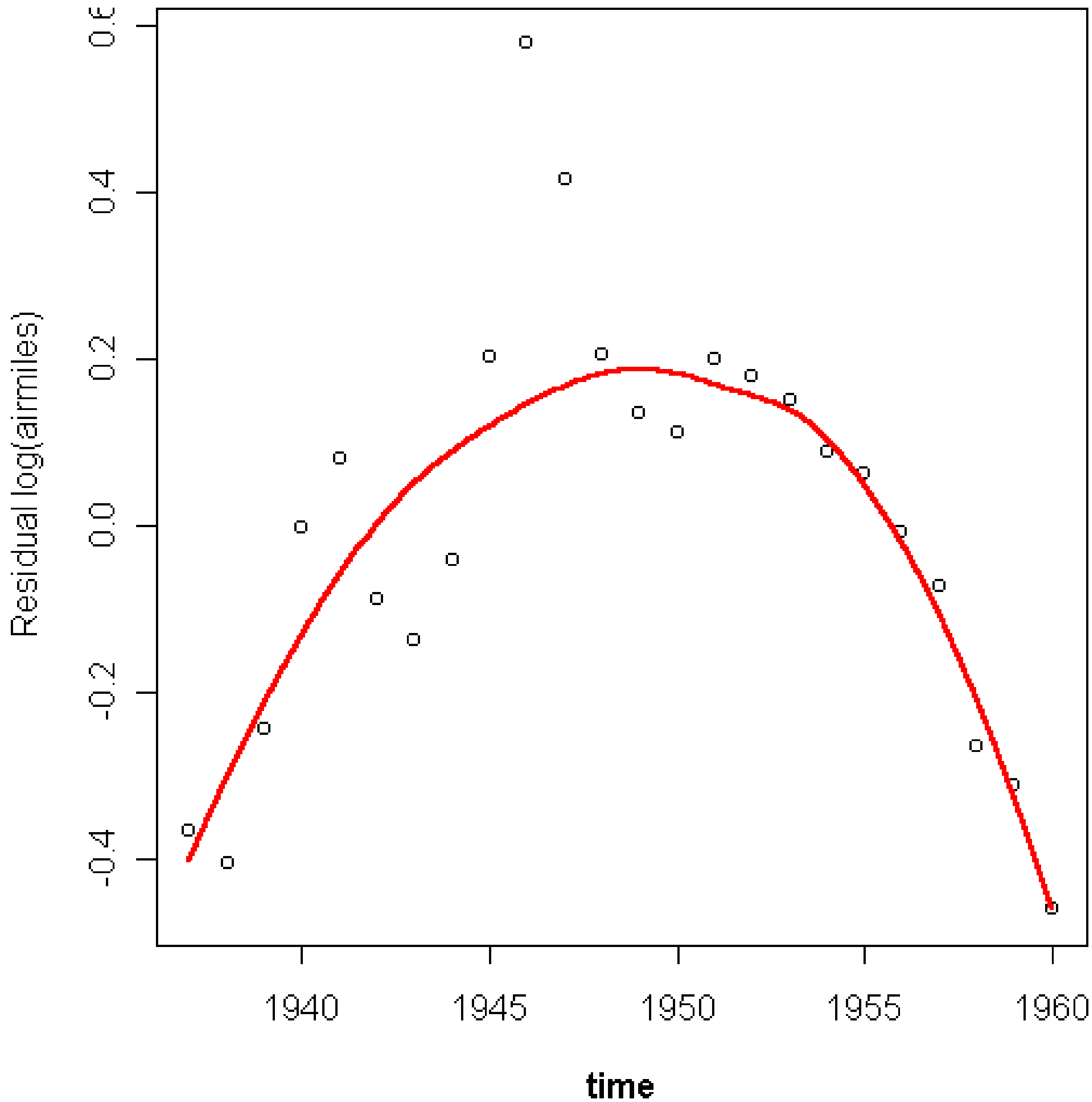
I(airmiles^(1/2))
50
100
150
1940
1945
1950
1955
1960
time


I(airmiles^(1/4))
12
10
8
6
1940
1945
1950
1955
1960
time

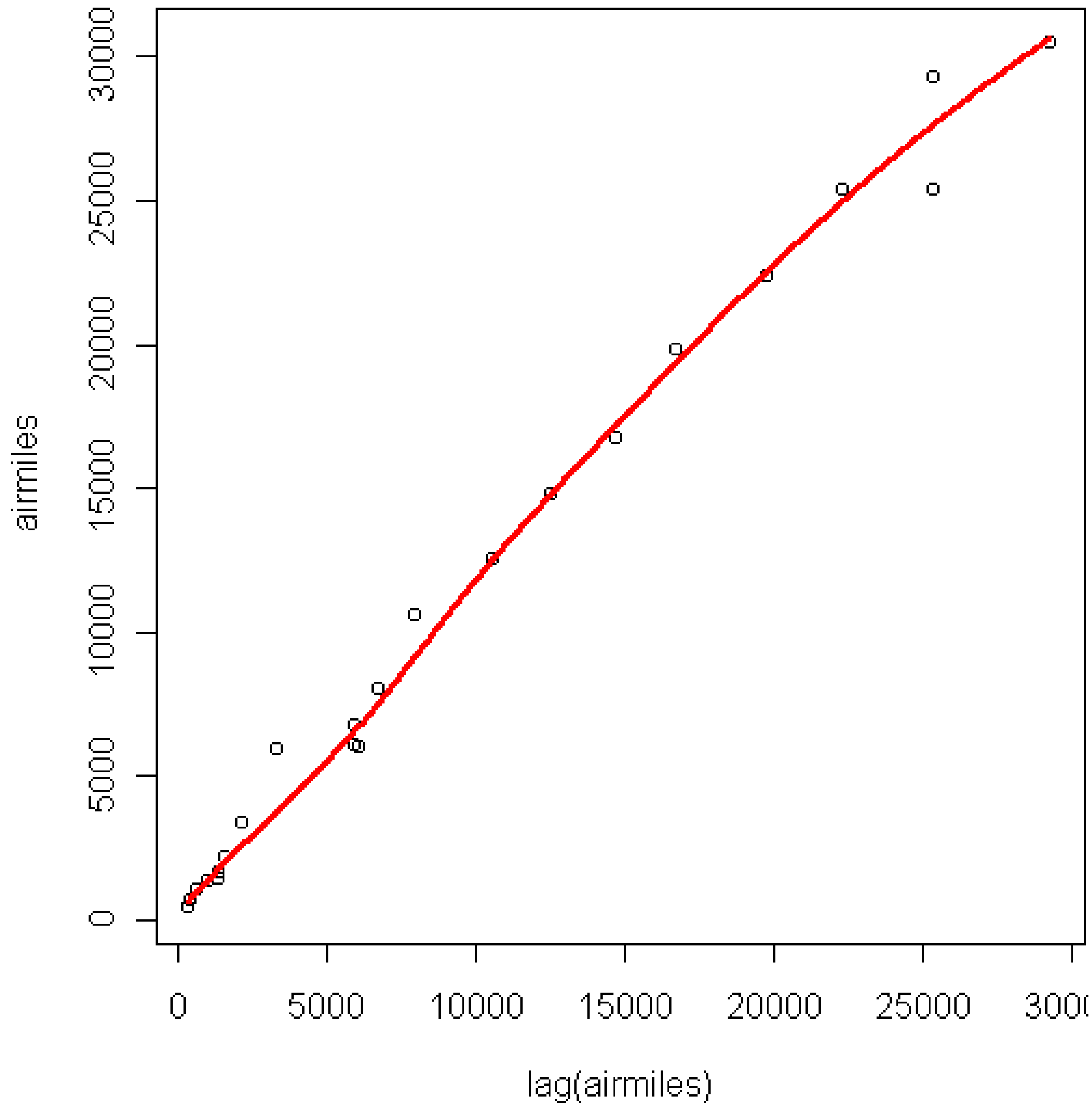


Residual I(airmiles^(1/2))
10
5
0
-5
-10
1940
1945
1950
1955
1960
time

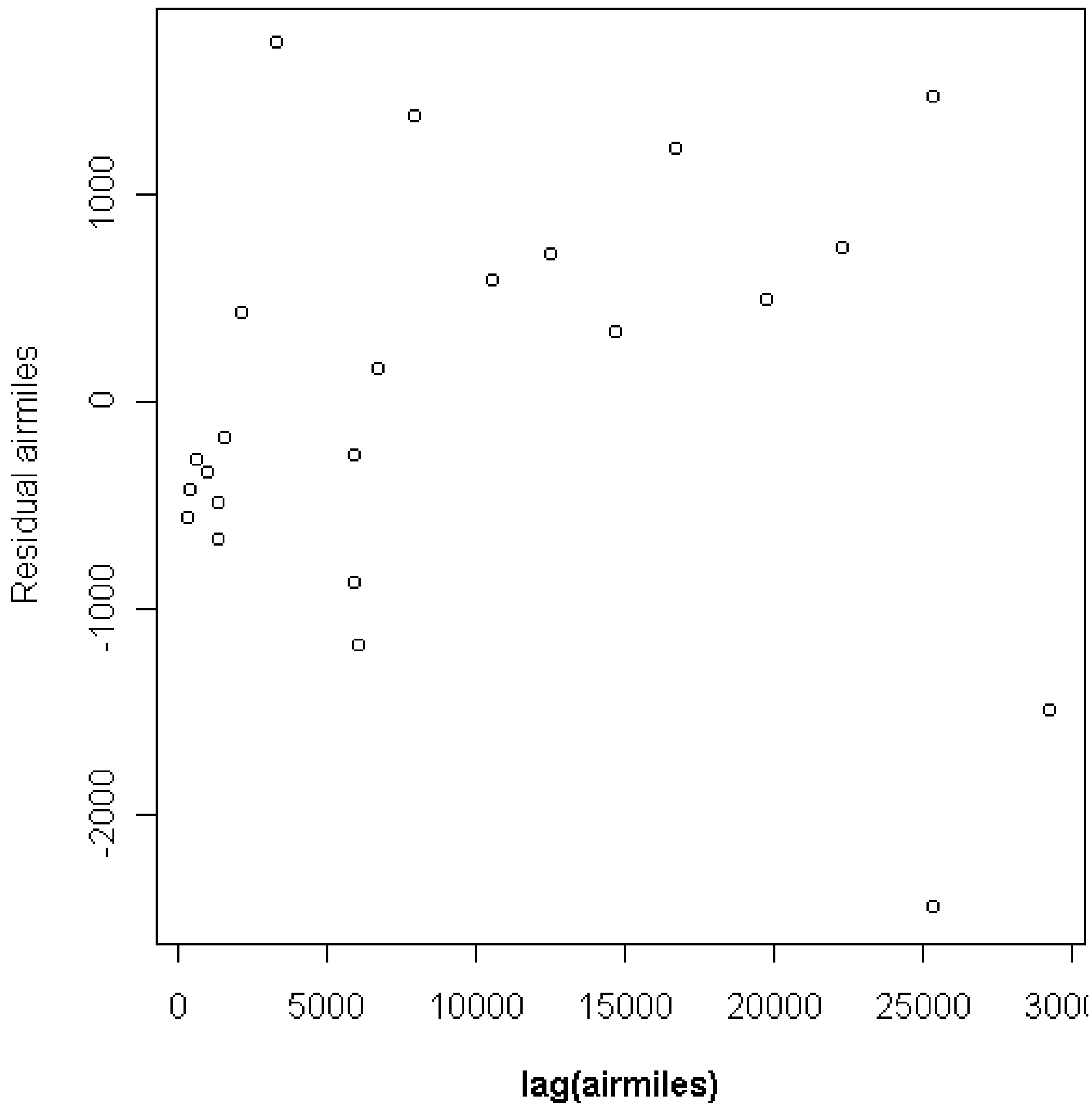

Residual I(airmiles^(1/4))
0.
0.6
0.4
0.2
0.0
-0.2
-0.4
1940
1945
1950
1955
1960
time

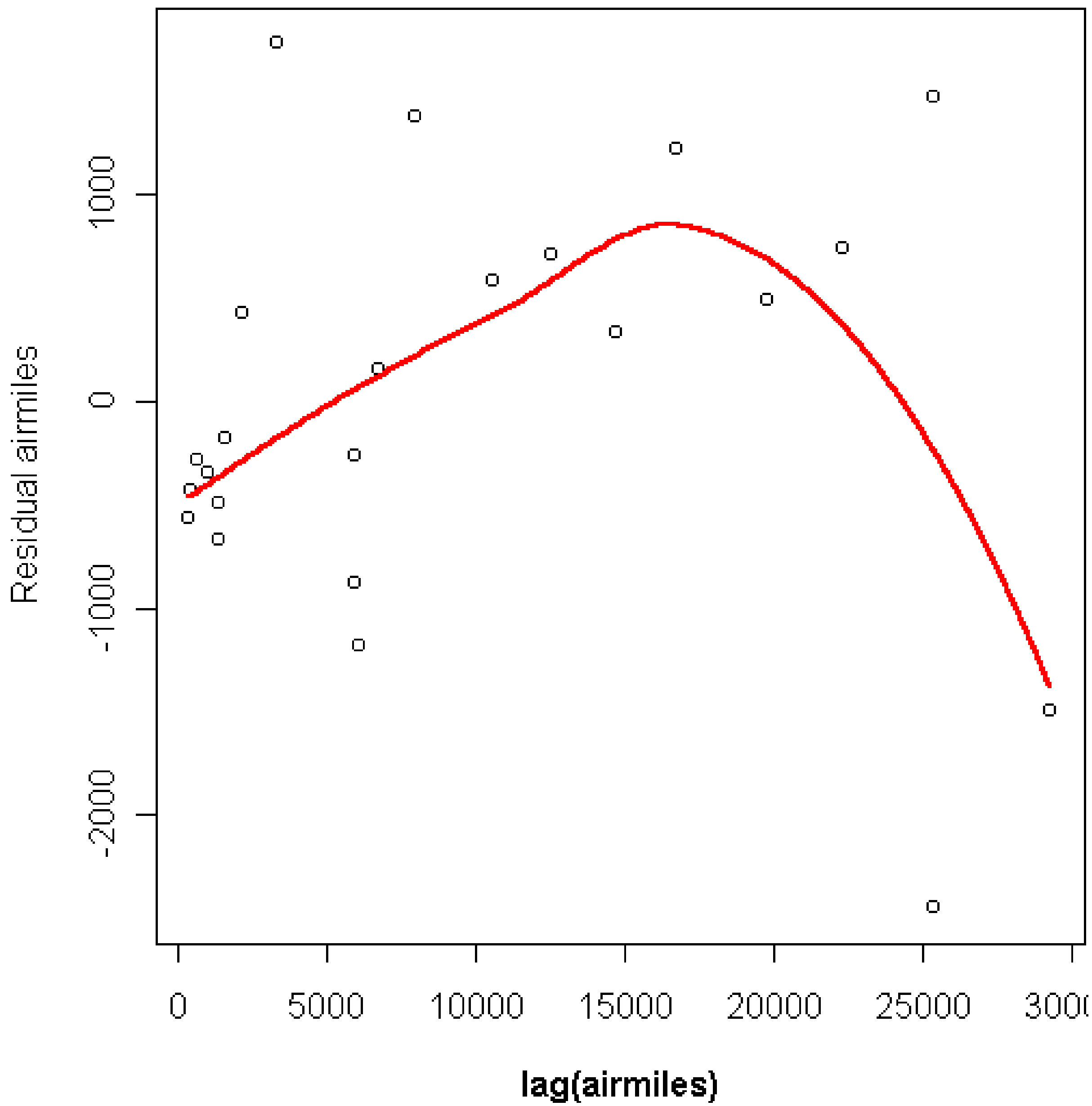


Residual log(airmiles)
0.6
0.4
0.2
0.0
-0.2
-0.4
1940
1945
1950
1955
1960
time

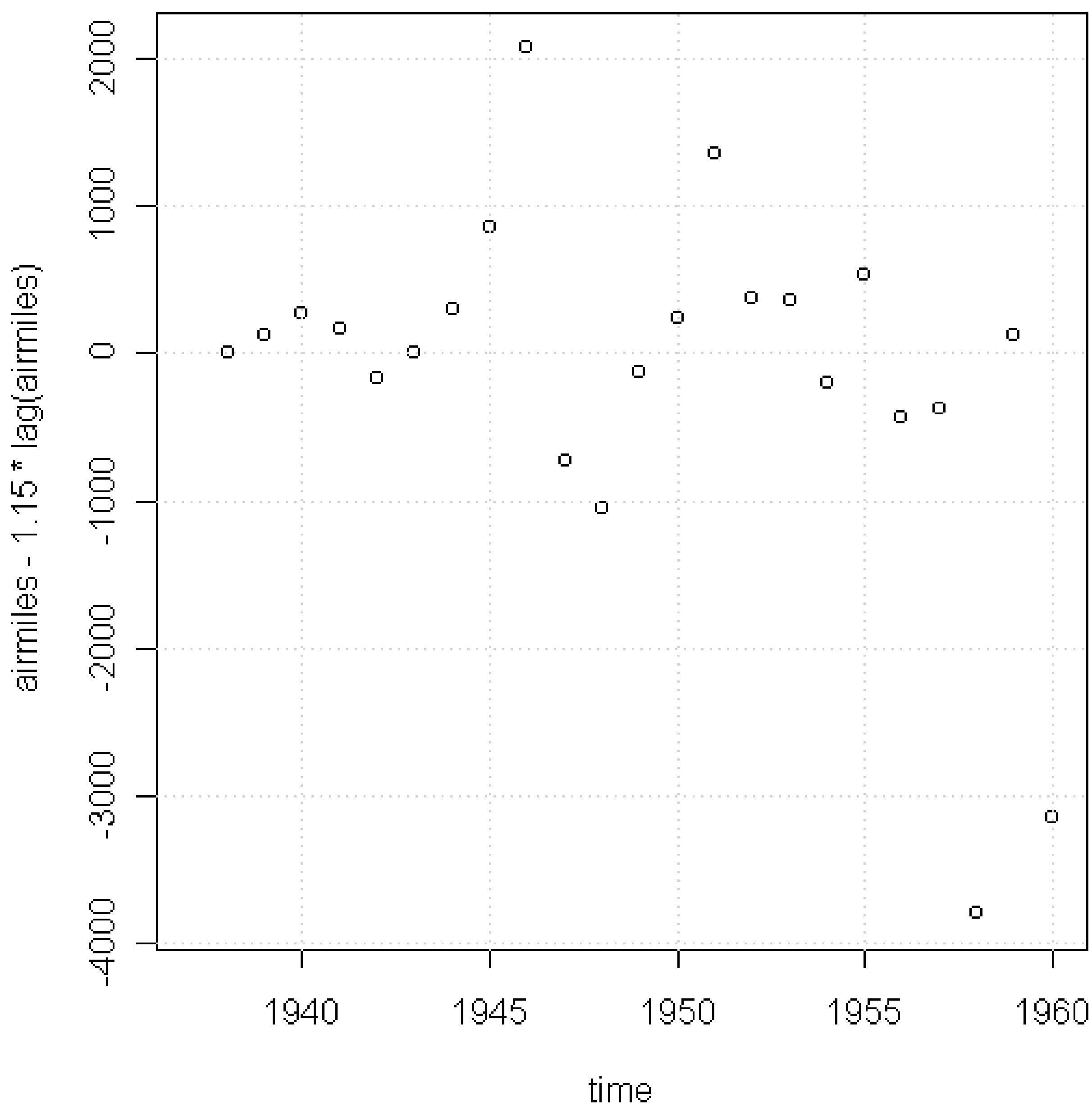


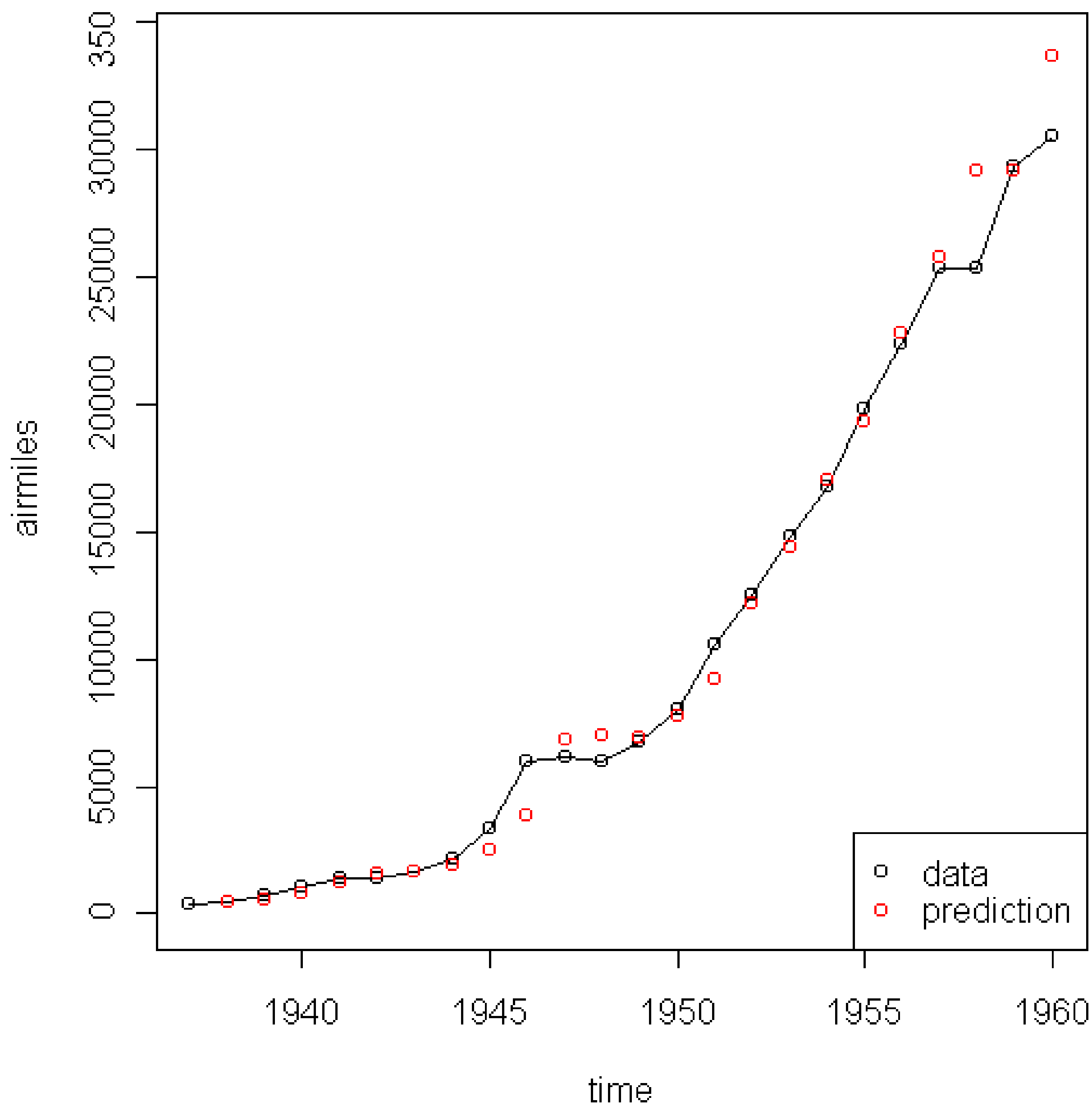
airmiles
0
5000
10000
15000
20000
25000
30000
lag(airmiles)


Residual airmiles
1000
0
-1000
-2000
0
5000
10000
15000
20000
25000
300(
lag(airmiles)



Residual airmiles
1000
0
-1000
-2000
0
5000
10000
15000
20000
25000
lag(airmiles)


airmiles - 1.15 * lag(airmiles)
2000
1000
0
-1000
-2000
-3000
-4000
1940
1945
1950
1955
1960
time


airmiles
0
5000
10000
15000
20000
25000
30000
350
time
1940
1945
1950
1955
1960
data
prediction

High-dimensional data

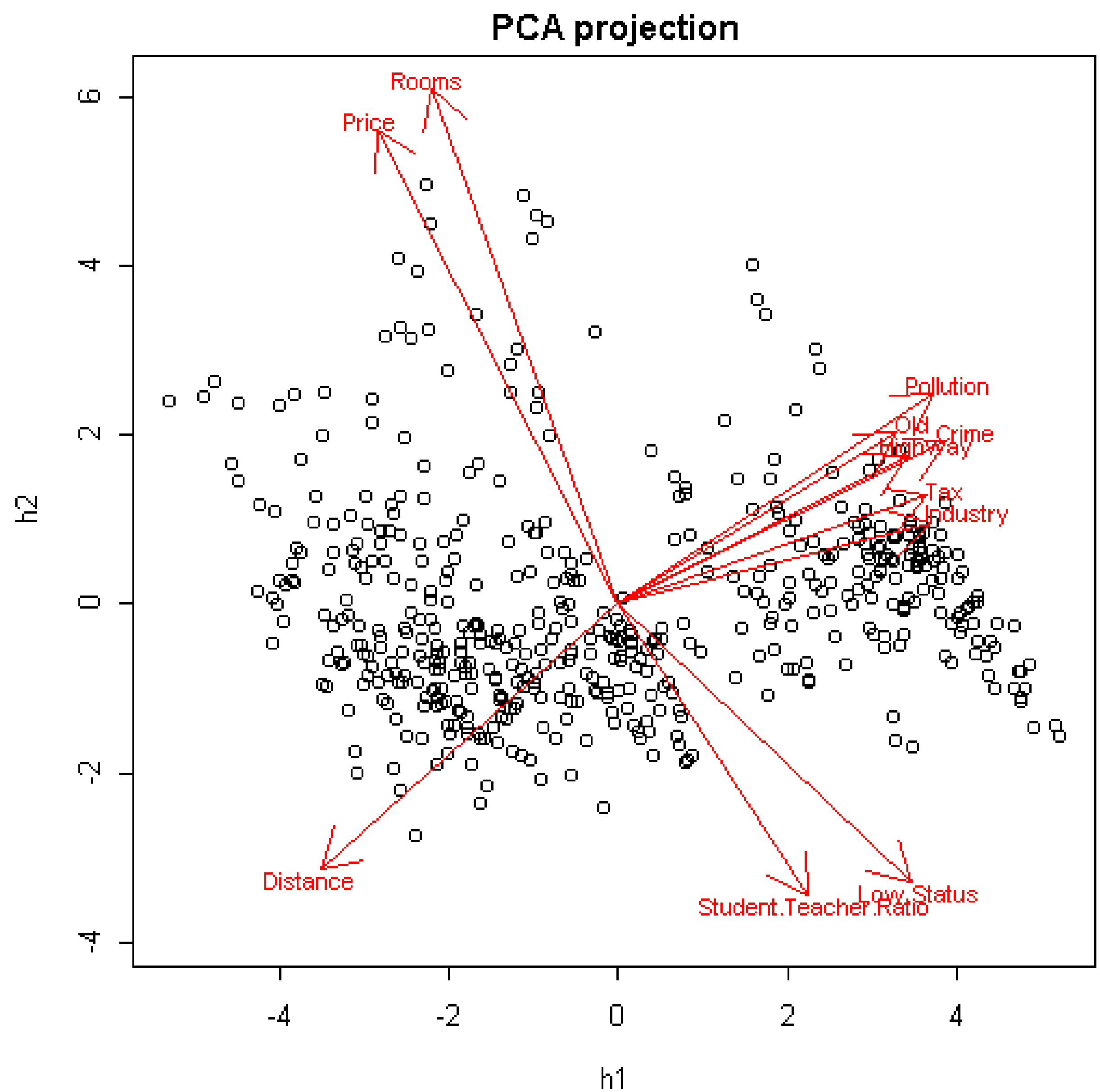
- Two basic approaches to visualization
- Many points, few dimensions:
  - Projection
  - Slicing
- Few points, many dimensions:
  - Parallel-coordinates
  - Iconic displays

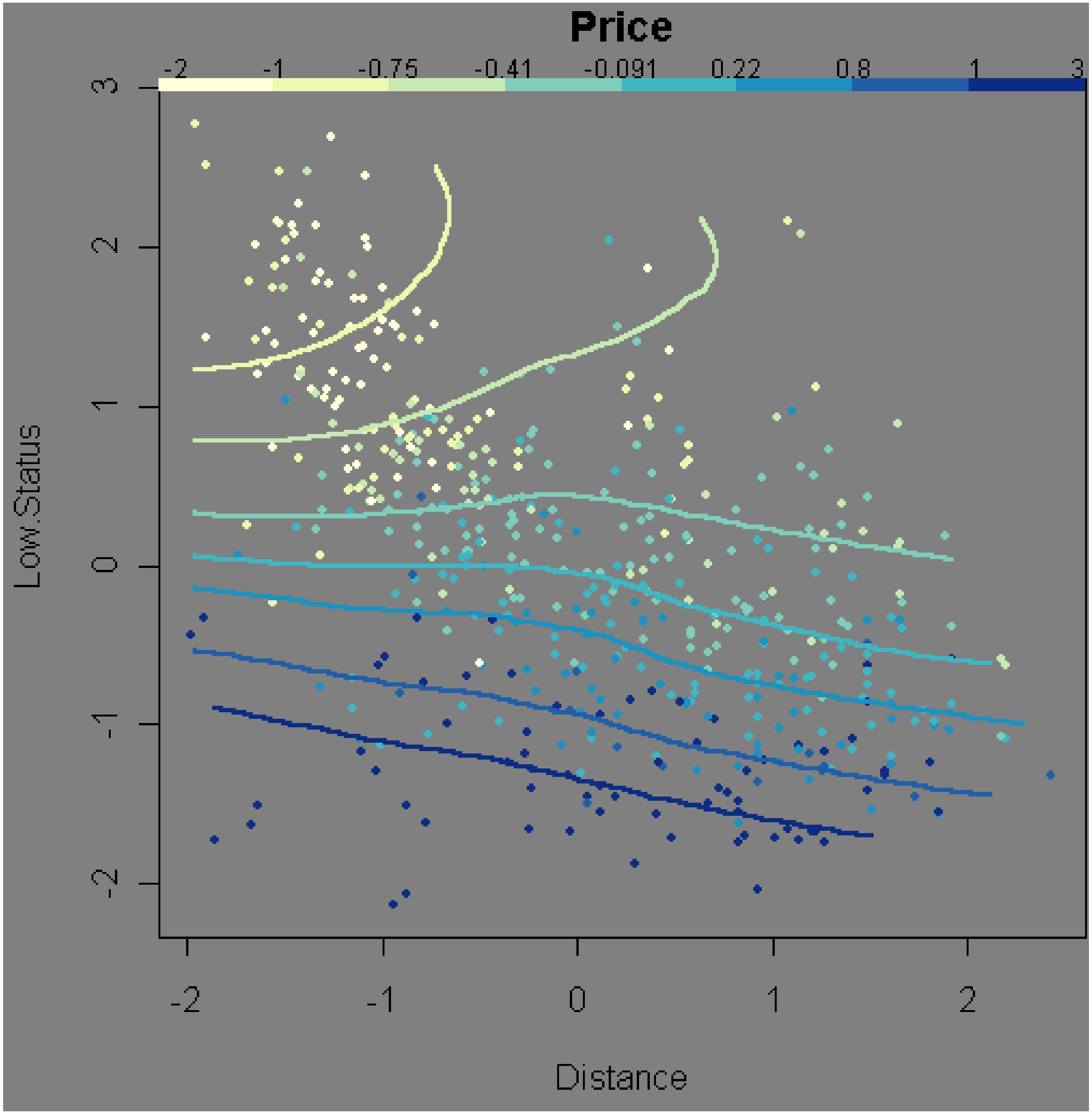
Projections

- PCA
  - Maximize the spread of the projected data
- Regression projection
  - Project only the predictors (inputs)
  - Maximize the spread of the response (output)

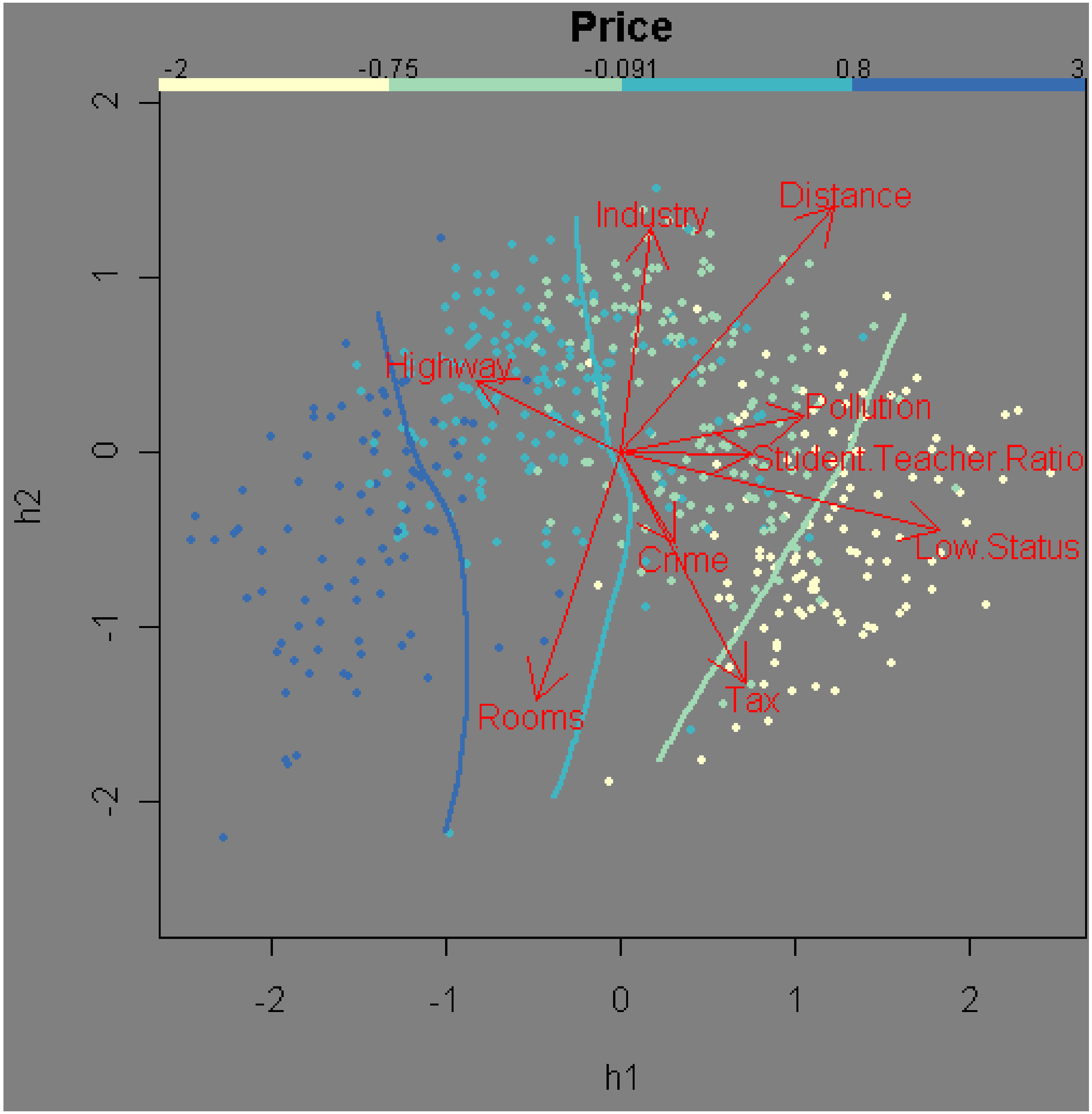
Boston Housing data

- Predict the median house value in Boston census tracts, based on crime, poverty, industry, pollution, etc.
- A regression problem with many predictors
- Is an additive model appropriate?


PCA projection
Rooms
Price
Pollution
Old
Crime
Highway
Tax
Industry
Distance
Student.Teacher.Ratio
Low.Status
h1
h2
6
4
2
0
-2
-4
-4
-2
0
2
4



Price
-2
-1
-0.75
-0.41
-0.091
0.22
0.8
1
3
Low.Status
Distance

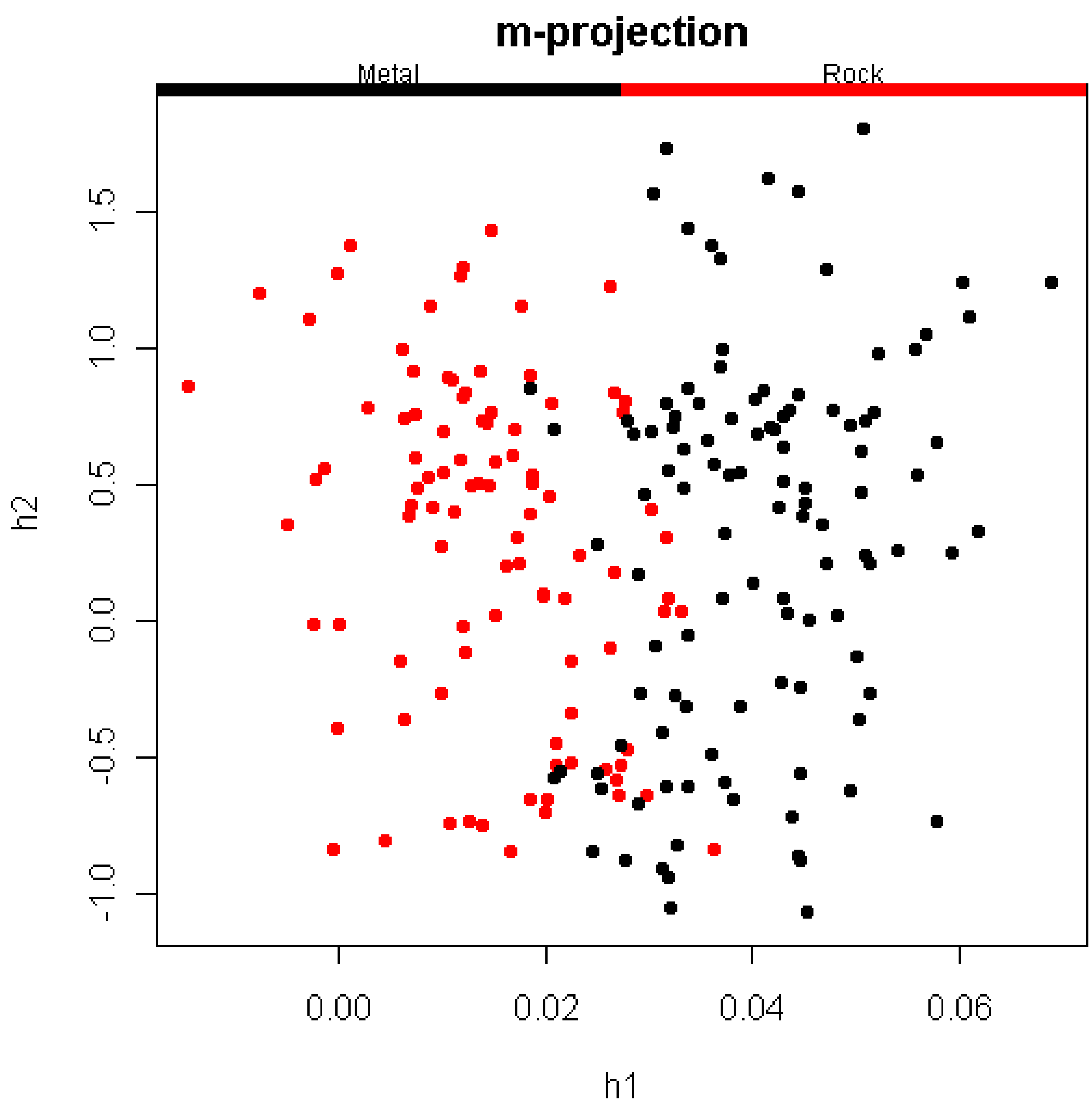

Price
-2
-0.75
-0.091
0.8
3
Highway
Industry
Distance
Pollution
Student.Teacher.Ratio
Low.Status
Crime
Tax
Rooms
h1
h2
-2
-1
0
1
2

Discriminative projections

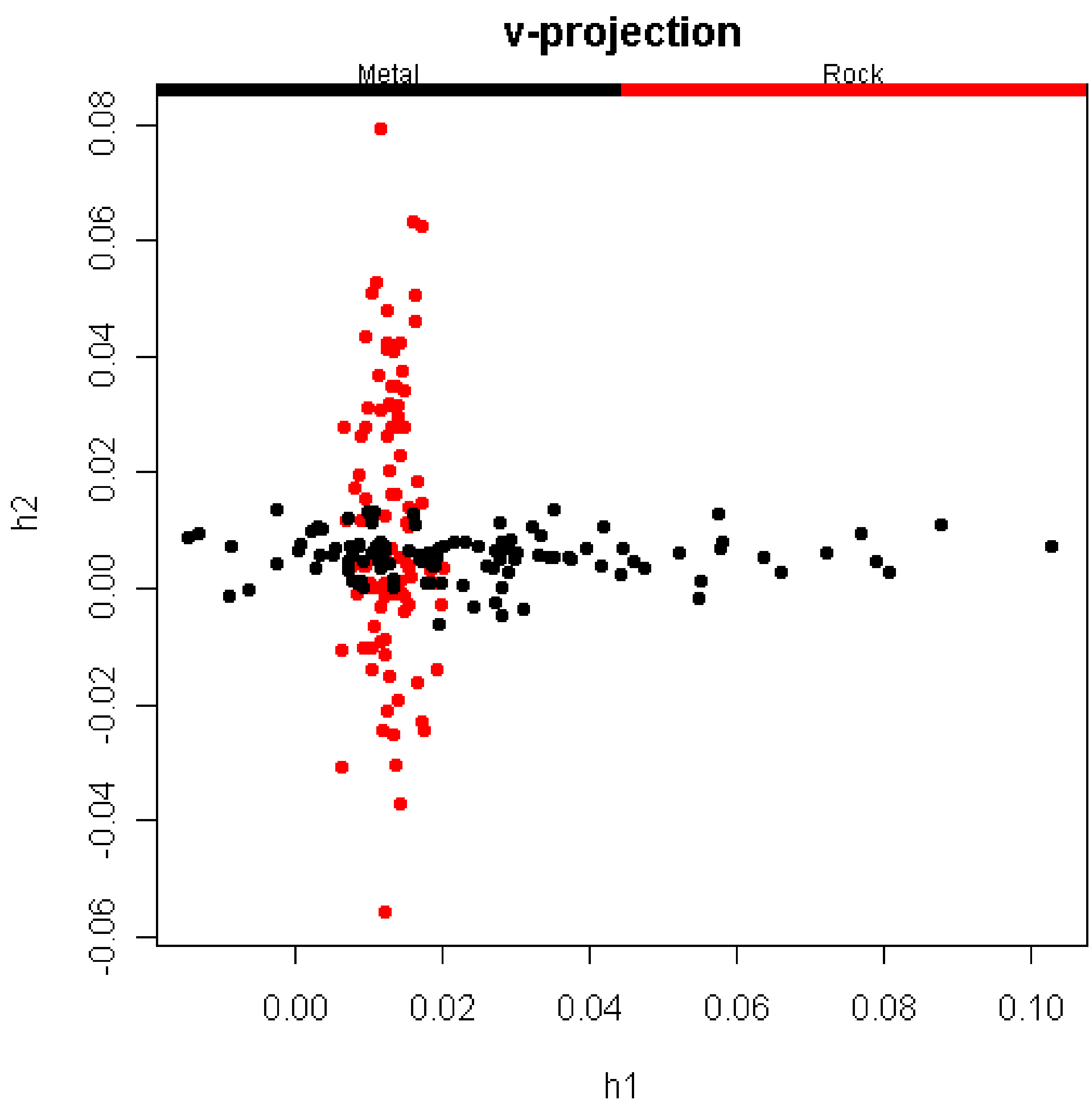
- M-projection (Fisher, LDA)
  - Tries to separate means of classes
- V-projection
  - Tries to separate variances of classes
- MV-projection
  - Maximize KL divergence between Gaussians
  - Separates means and variances

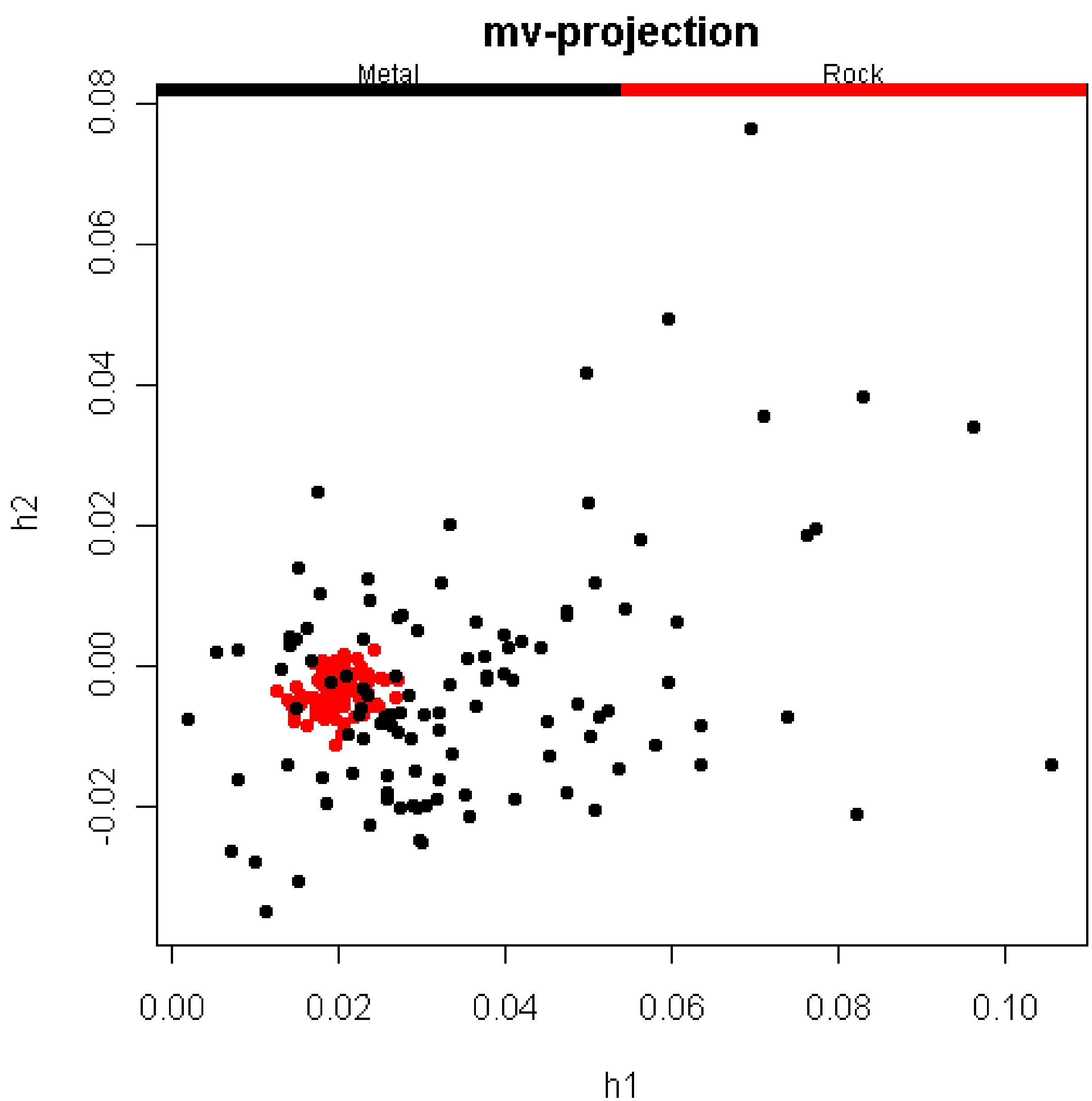
Sonar problem

- Sonar echo is represented by energy in 60 frequency bands
- Mines vs. Rocks
- Dataset is linearly separable, but 1nn consistently beats linear classifiers

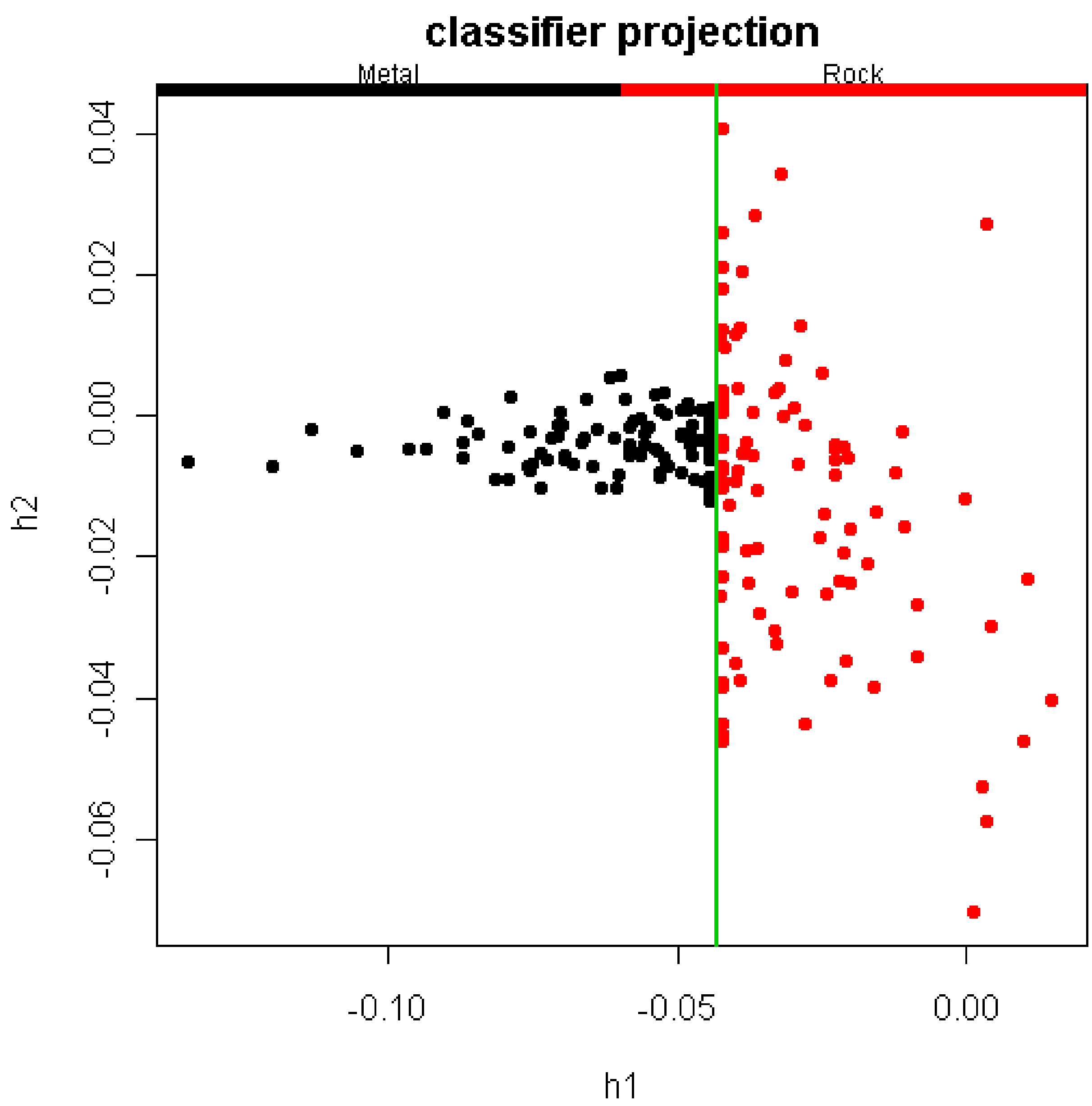


m-projection
Metal
Rock
h2
1.5
1.0
0.5
0.0
-0.5
-1.0
0.00
0.02
0.04
0.06
h1


v-projection
Metal
Rock
h2
h1
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
0.00
0.02
0.04
0.06
0.08
0.10

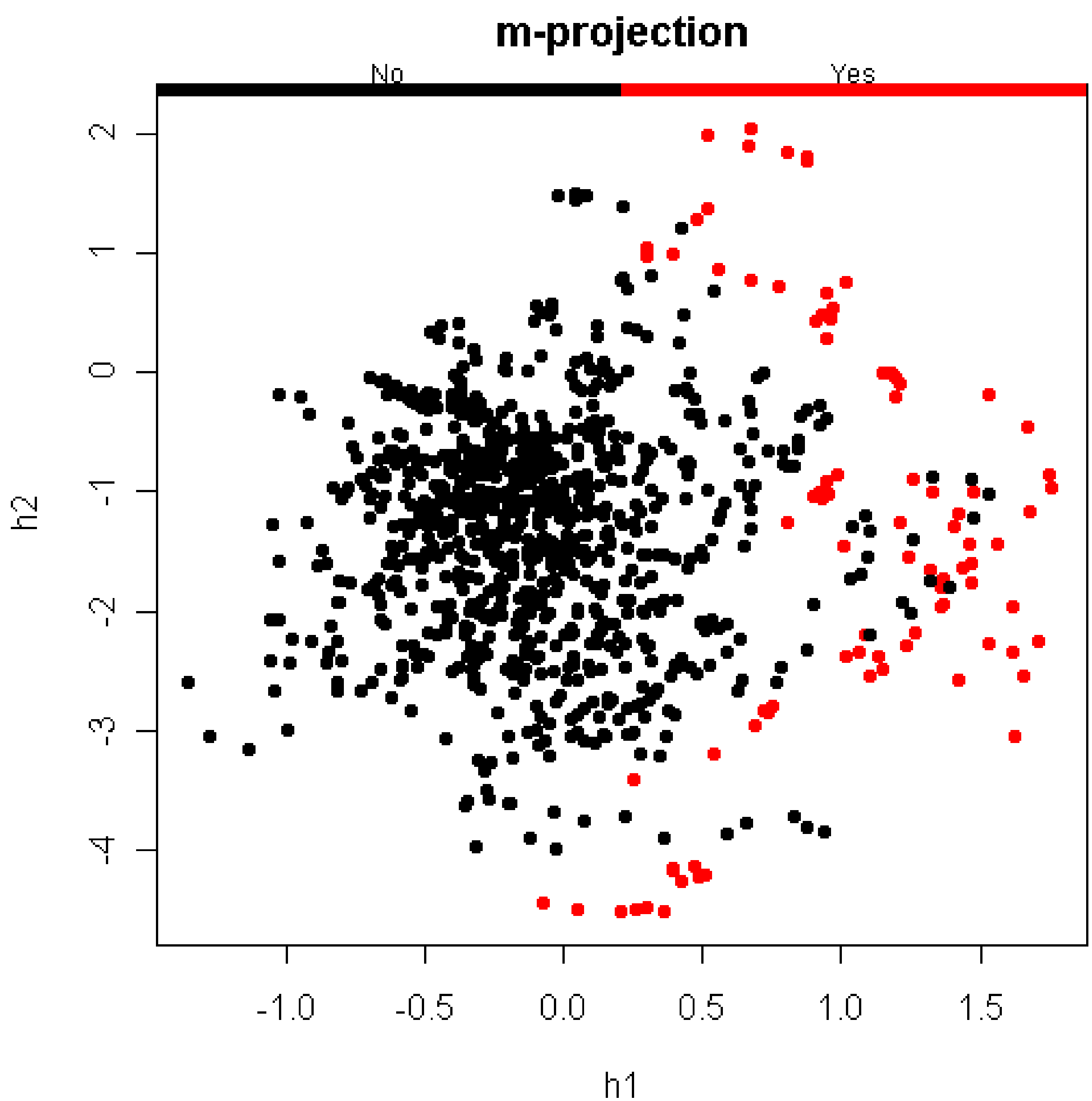


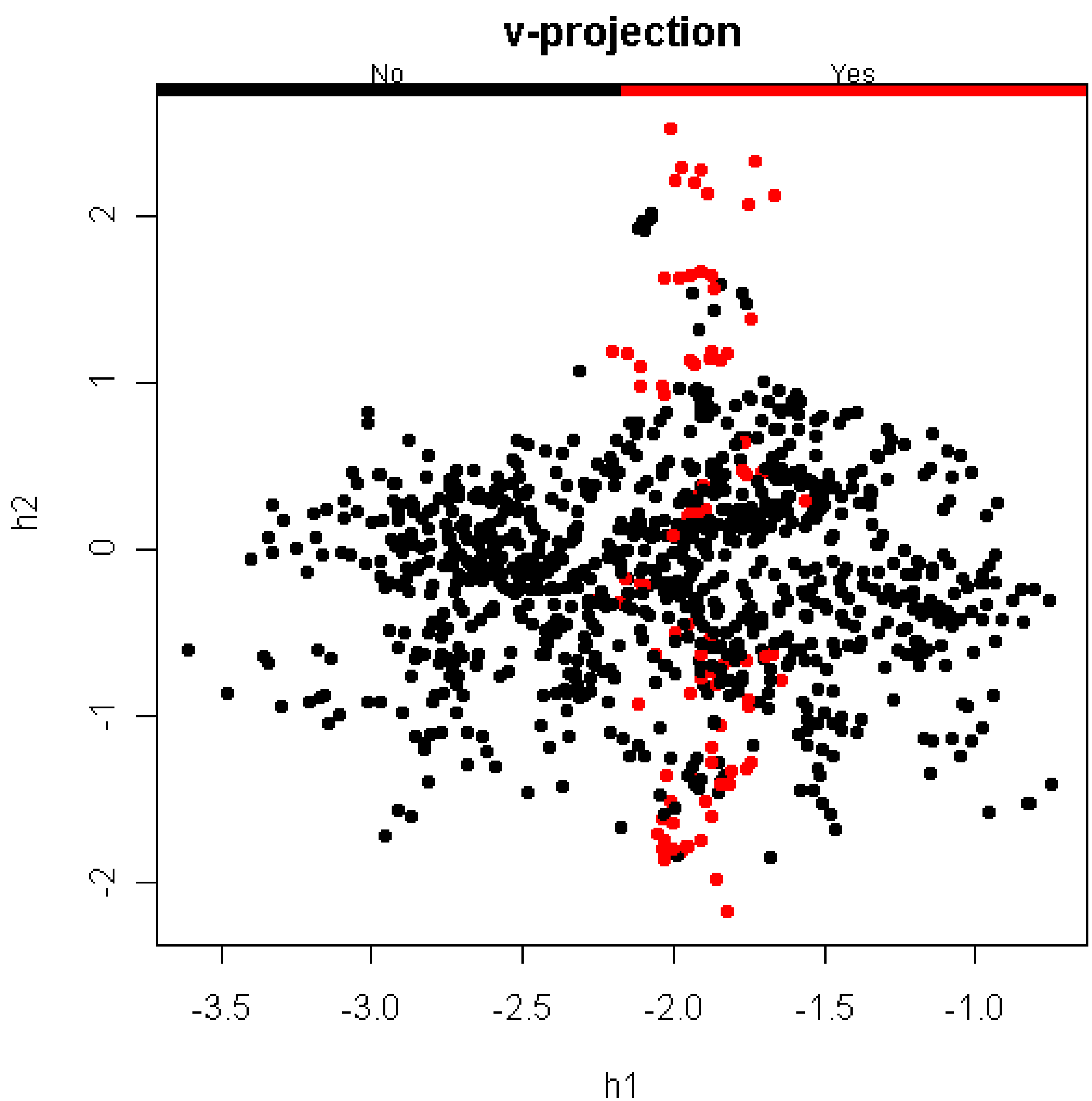
mv-projection
Metal
Rock
h2
h1
0.08
0.06
0.04
0.02
0.00
-0.02
0.00
0.02
0.04
0.06
0.08
0.10

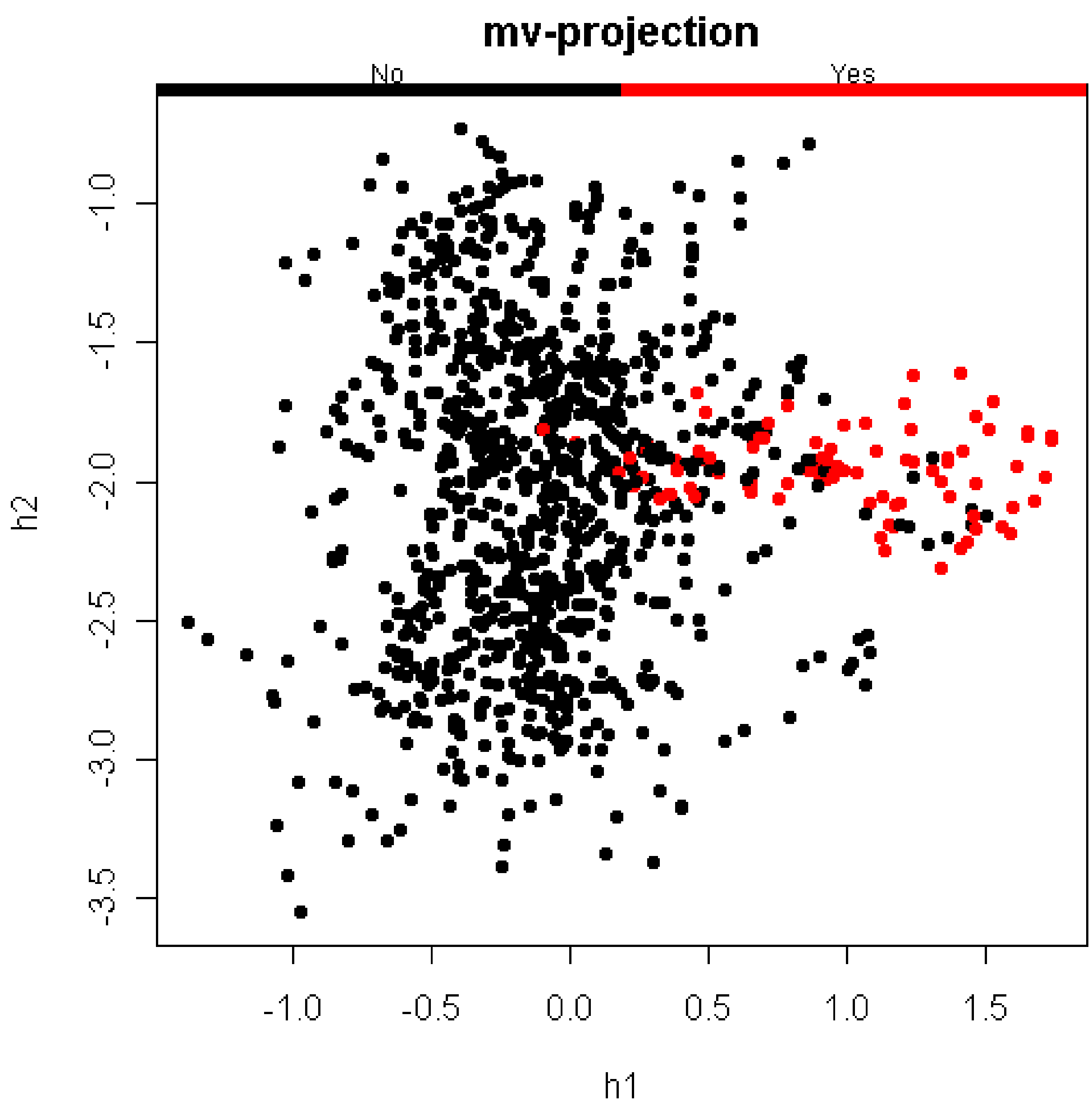

classifier projection
Metal
Rock
0.04
0.02
0.00
-0.02
-0.04
-0.06
h2
-0.10
-0.05
0.00
h1

Vowels dataset

- Binary problem: “hid” vs. rest
- Knn and quadratic kernel beat linear

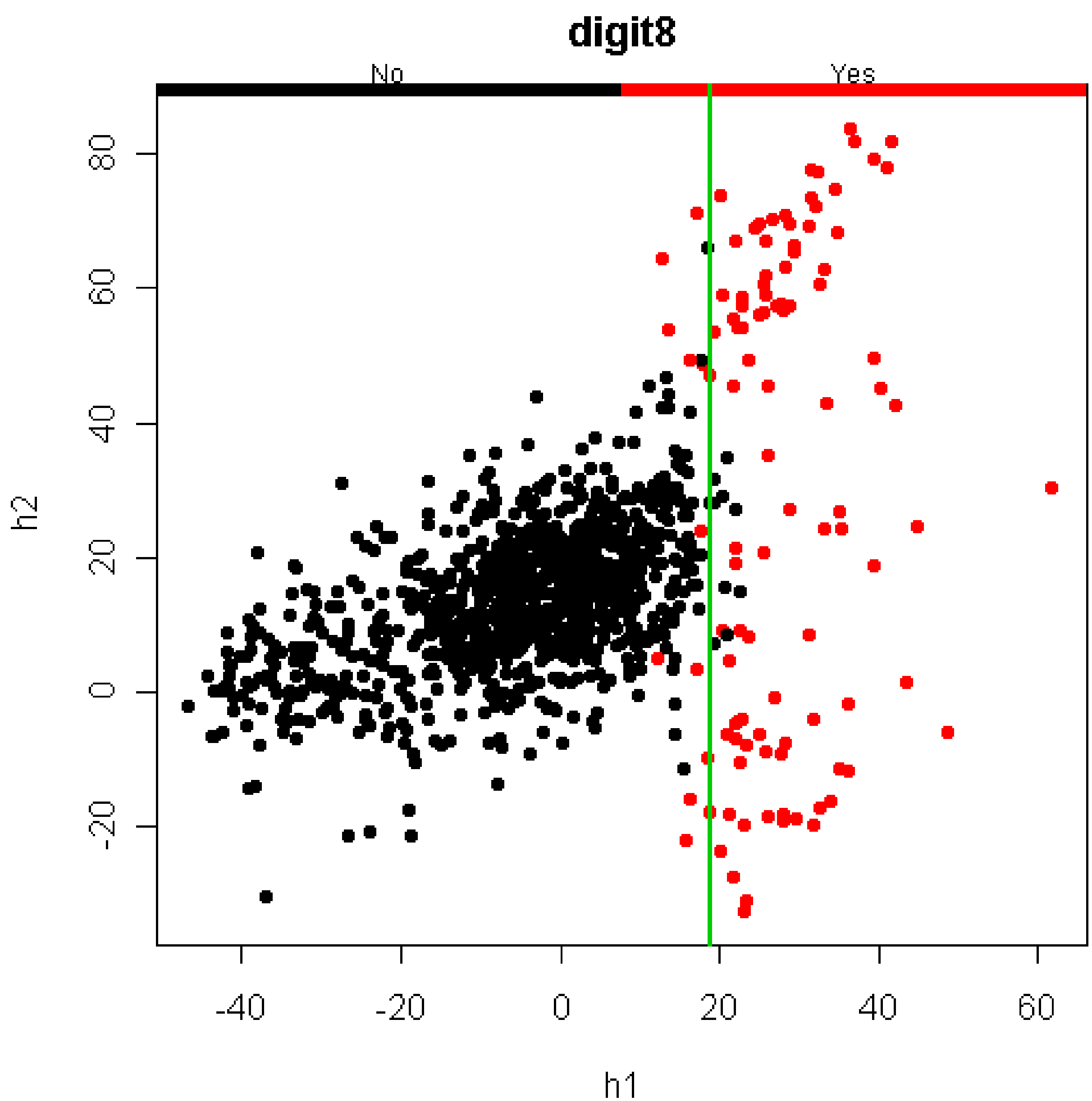

m-projection
No
Yes
2
1
0
-1
-2
-3
-4
h2
-1.0
-0.5
0.0
0.5
1.0
1.5
h1

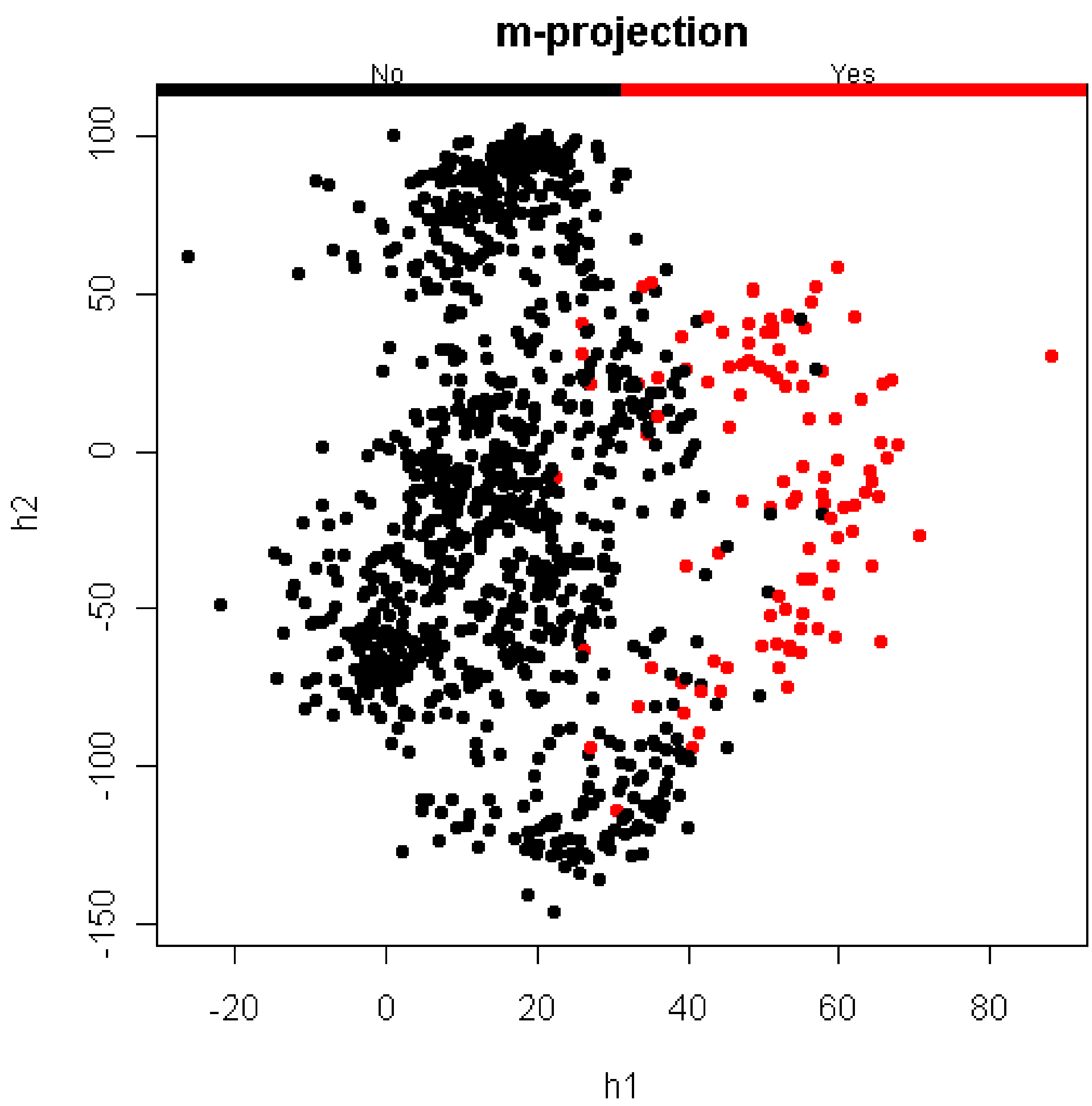

v-projection
No
Yes
h2
2
1
0
-1
-2
-3.5
-3.0
-2.5
-2.0
-1.5
-1.0
h1

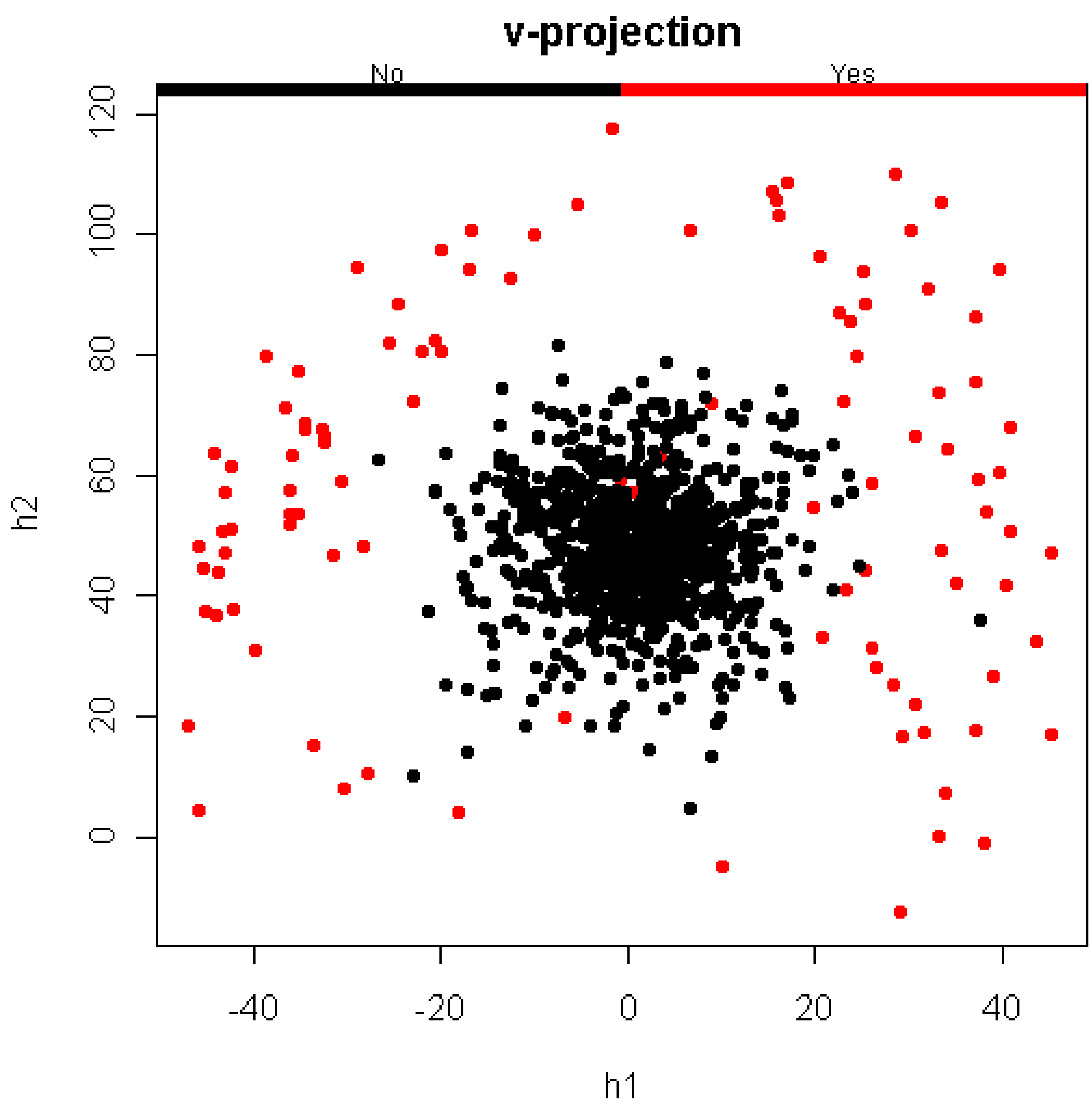

mv-projection
No
Yes
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
h2
-1.0
-0.5
0.0
0.5
1.0
1.5
h1

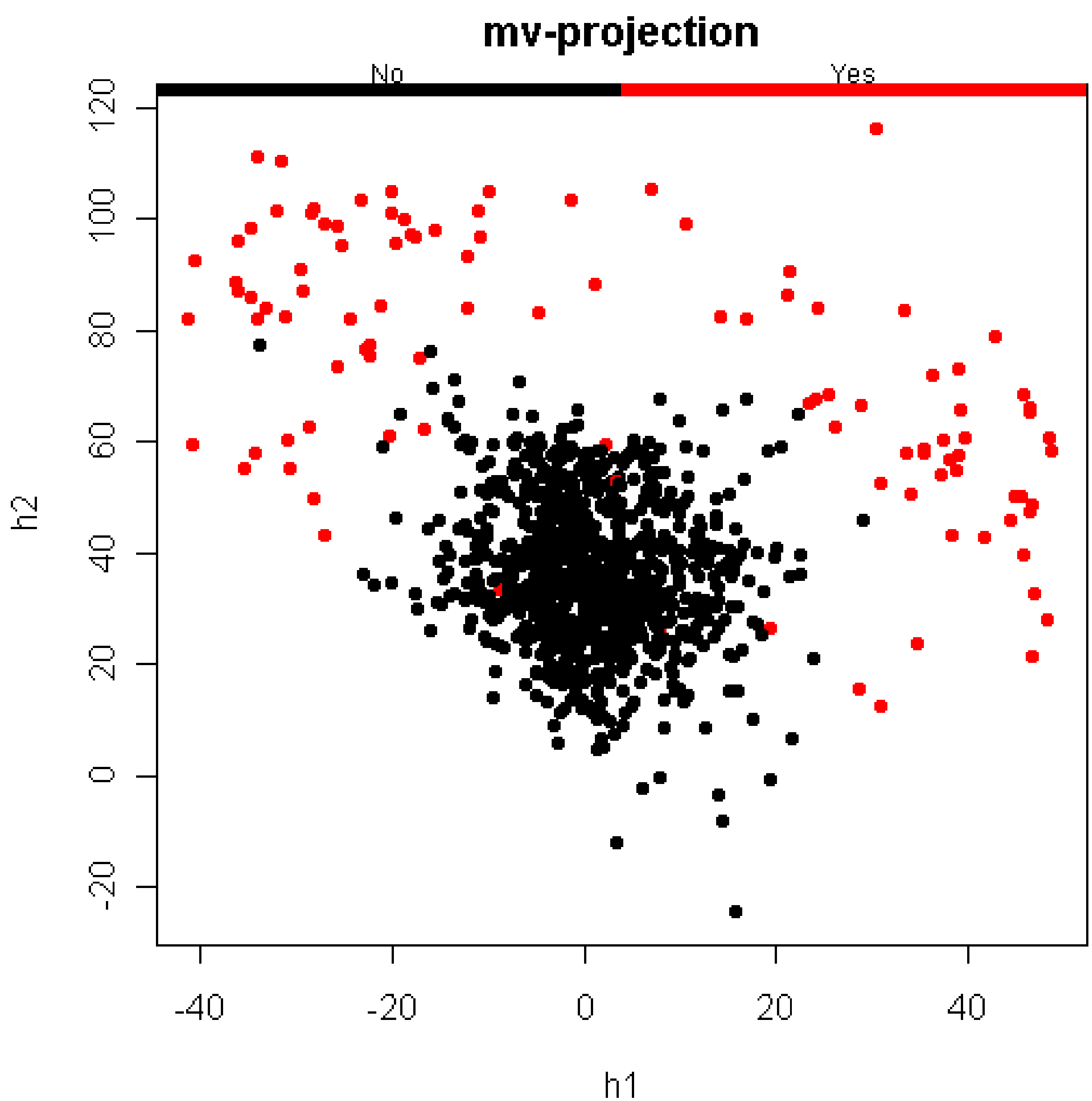
Online digit recognition

- Classify “8” vs. rest
- Knn beats quadratic kernel beats linear


digit8
No
Yes
80
60
40
20
0
-20
h2
-40
-20
0
20
40
60
h1


m-projection
No
Yes
h2
100
50
0
-50
-100
-150
-20
0
20
40
60
80
h1


v-projection
No
Yes
h2
120
100
80
60
40
20
0
h1
-40
-20
0
20
40


mv-projection
No
Yes
120
100
80
60
40
20
0
-20
h2
-40
-20
0
20
40
h1

Some good books

- “The Elements of Graphing Data”, William Cleveland, 2nd Ed.
- “Visualizing Data”, William Cleveland
- “The Visual Display of Quantitative Information”, Edward Tufte
- “Exploratory Data Analysis”, John Tukey

Summary

- Visualization is a simple and fast way to check model assumptions and learn about a domain
- Many opportunities still exist to design better graphs, esp. for high dimensions
- Visualization is not "art", but a well - structured field, worthy of research attention