Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Smoothing

Sometimes we want to smooth relations

Tukey phrased this as

- data = smooth + rough
- The smoothed version should show patterns not evident in raw data
- The rough should have no systematic variation

Many of these methods are nonparametric

Some are parametric

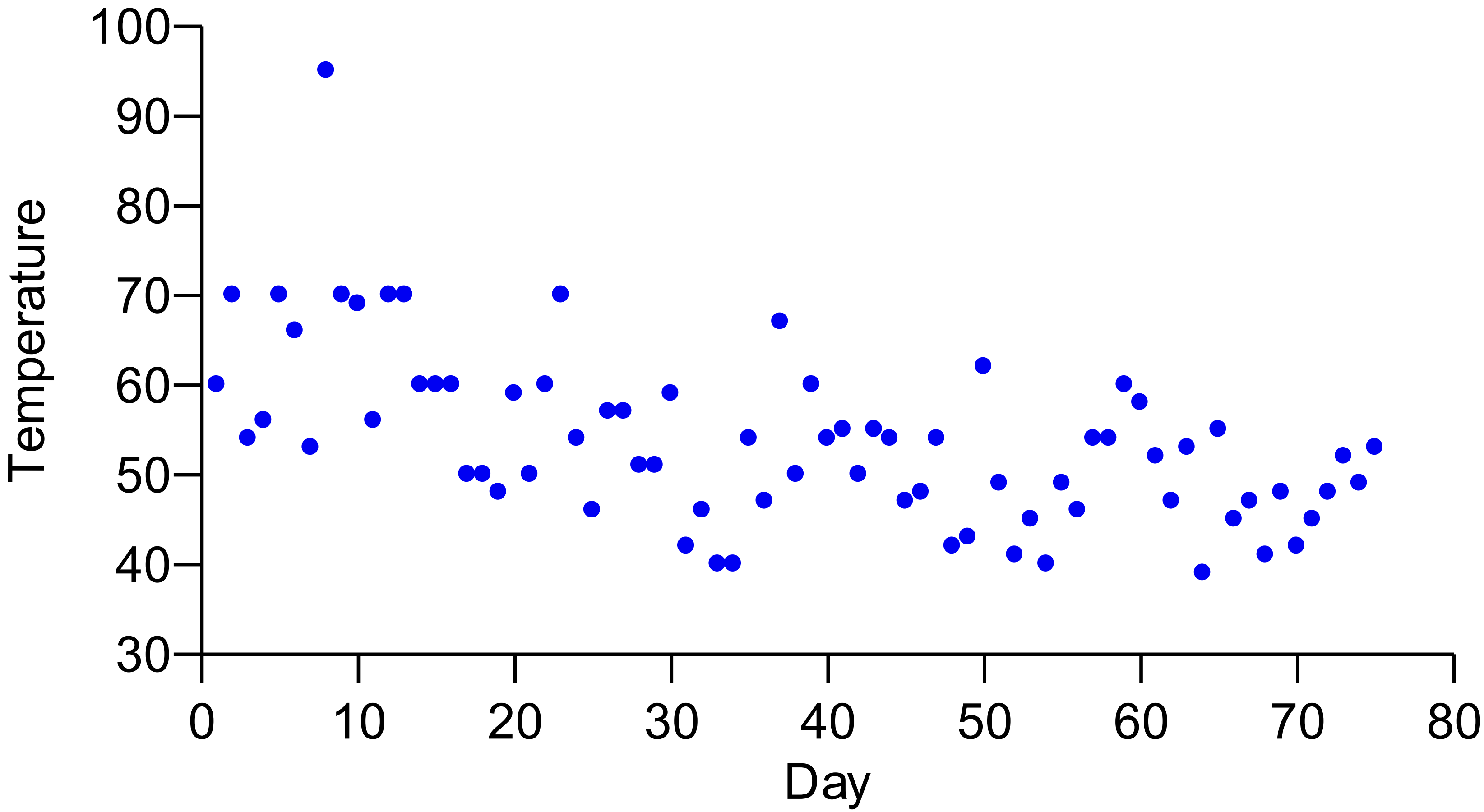
We use them to discover, not to confirm

Stephen Stigler (Seven pillars of statistical wisdom, JSM 2014)

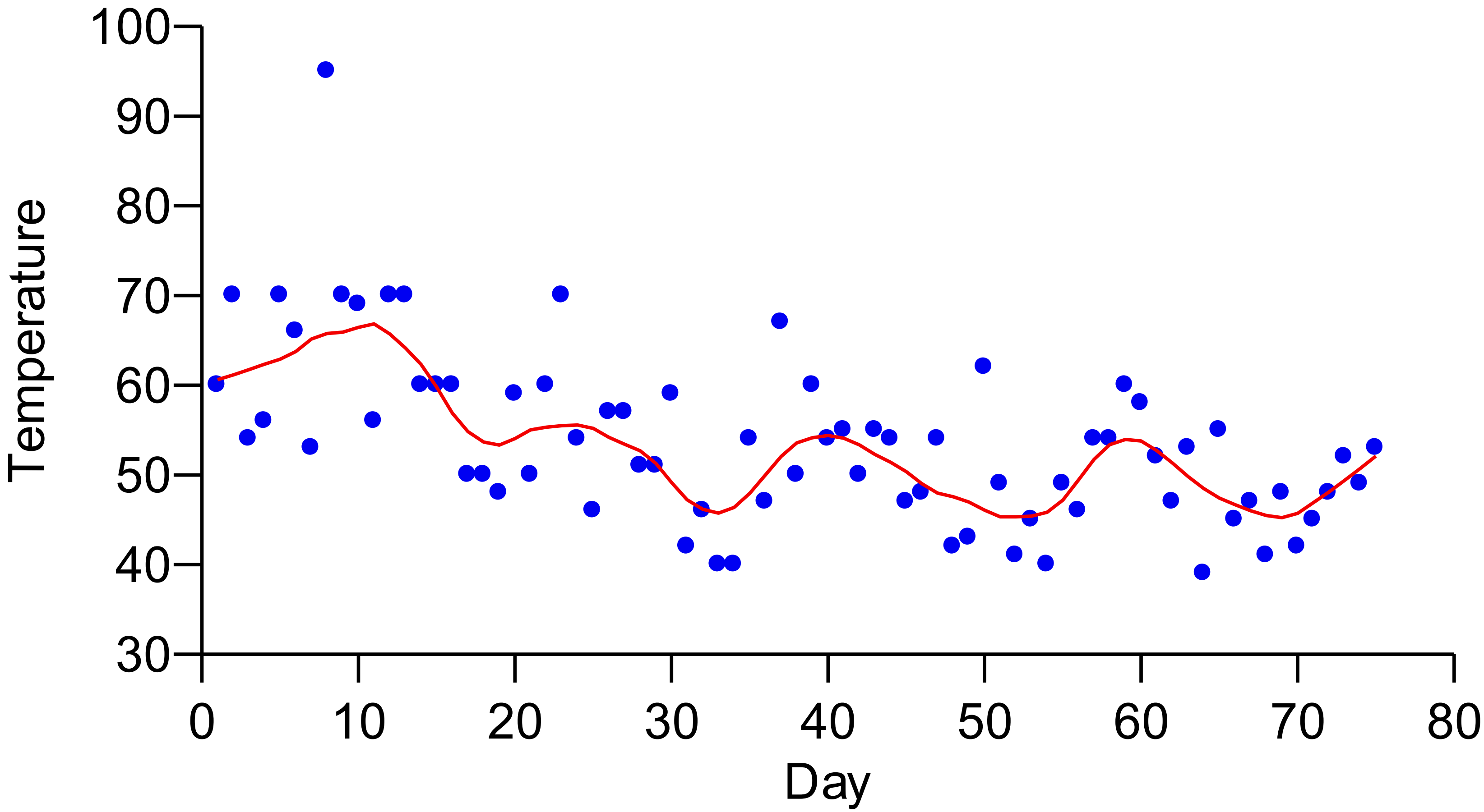
Fallacy: Discarding individual level data reduces the amount of information.

Truth: Discarding individual level data, by aggregating or averaging, can increase information.

Smoothing

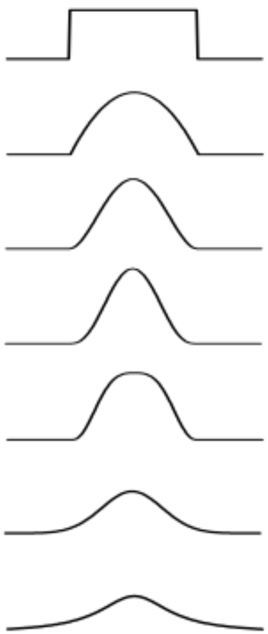


Smoothing



Smoothing

Smoothing windows (kernels)



uniform: $f(x) = a : (-w \le x \le w)$, else 0

epanechnikov: $f(x) = a(1 - (x/w))^2 : (-w \le x \le w)$, else 0

biweight: $f(x) = a(1 - (x/w)^2)^2 : (-w \le x \le w)$, else 0

triweight: $f(x) = a(1 - (x/w)^2)^3 : (-w \le x \le w)$, else 0

tricube: $f(x) = a(1 - |x/w|^3)^3 : (-w \le x \le w)$, else 0

gaussian: $f(x) = ae^{-(x/w)^2}$

cauchy: $f(x) = a/(b + (x/w)^2)$

Smoothing

Smoothing Functions

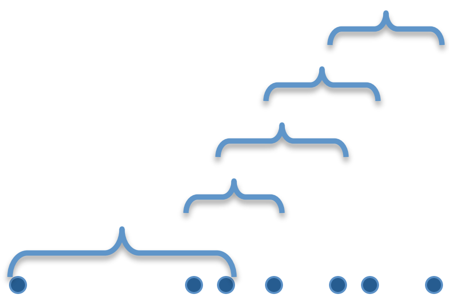
- Kernel smoothing
  - mean
  - median
  - mode
- Polynomial smoothing
  - linear regression
  - quadratic regression
  - etc.

Smoothing

Bandwidth

Neighborhood

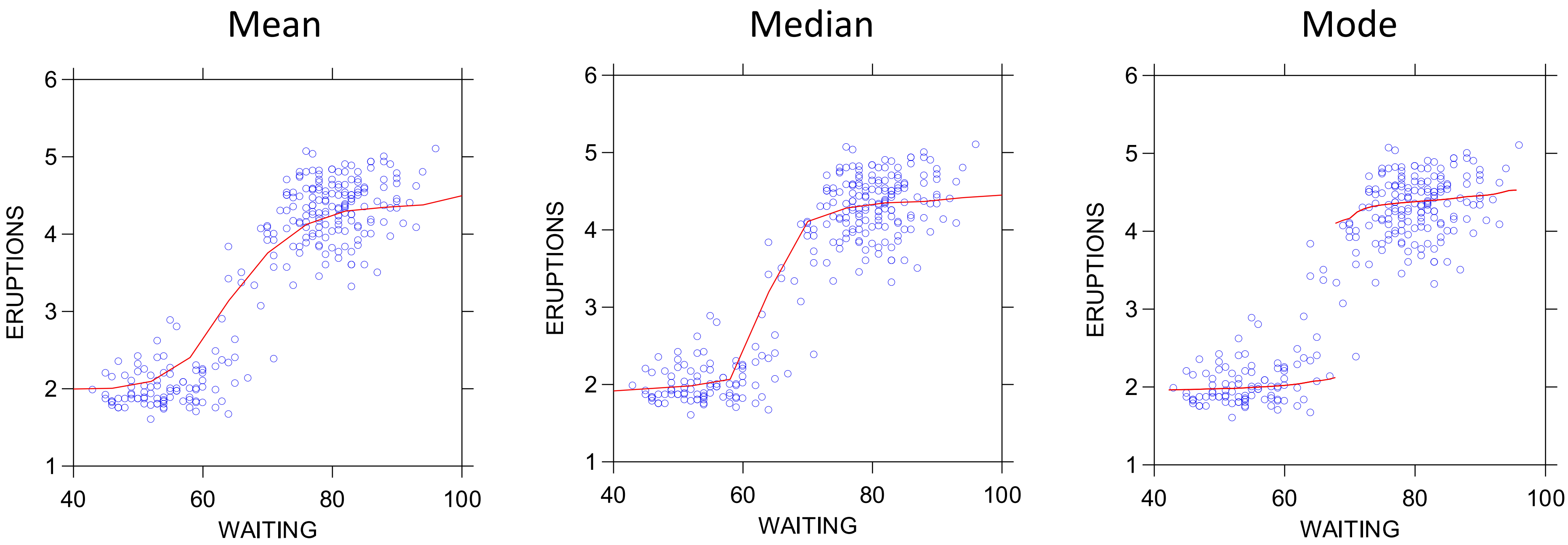
k-nearest neighbor (KNN)



fixed bandwidth

Smoothing

Kernel smoothers

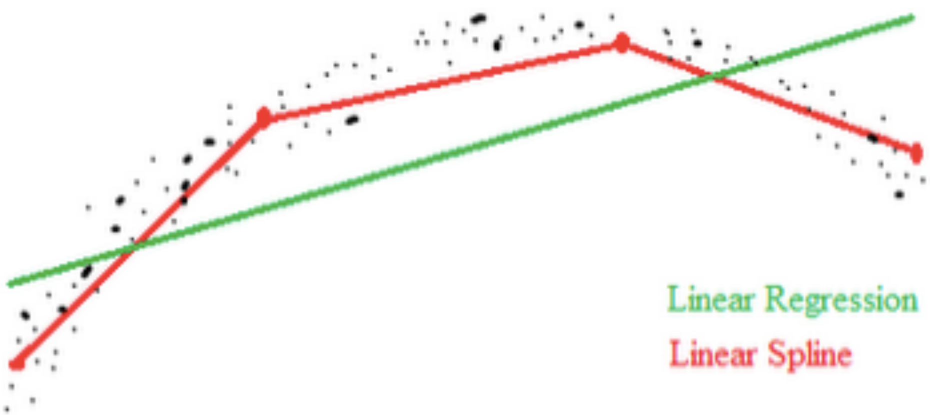


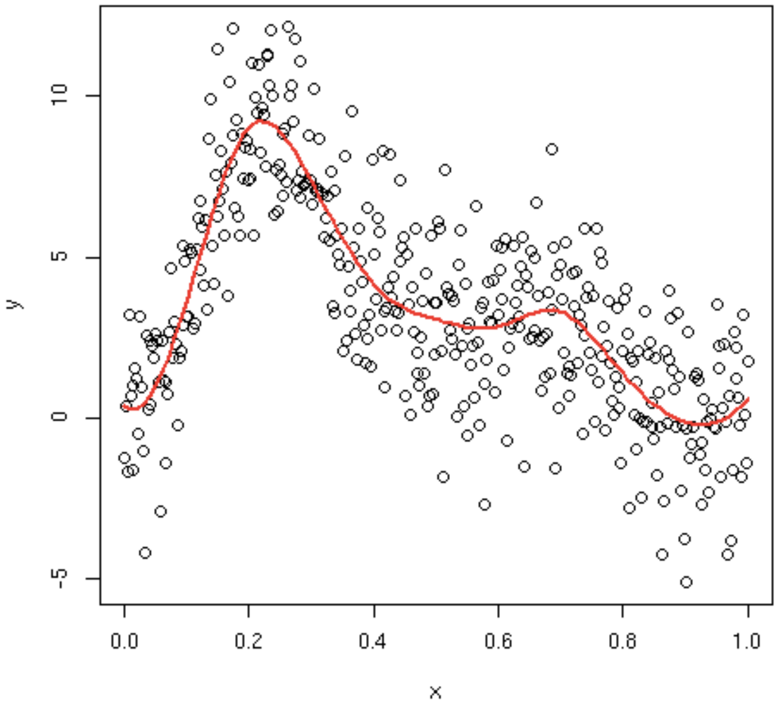
Smoothing

Polynomial Smoothers

Spline Regression

Piecewise polynomial regression

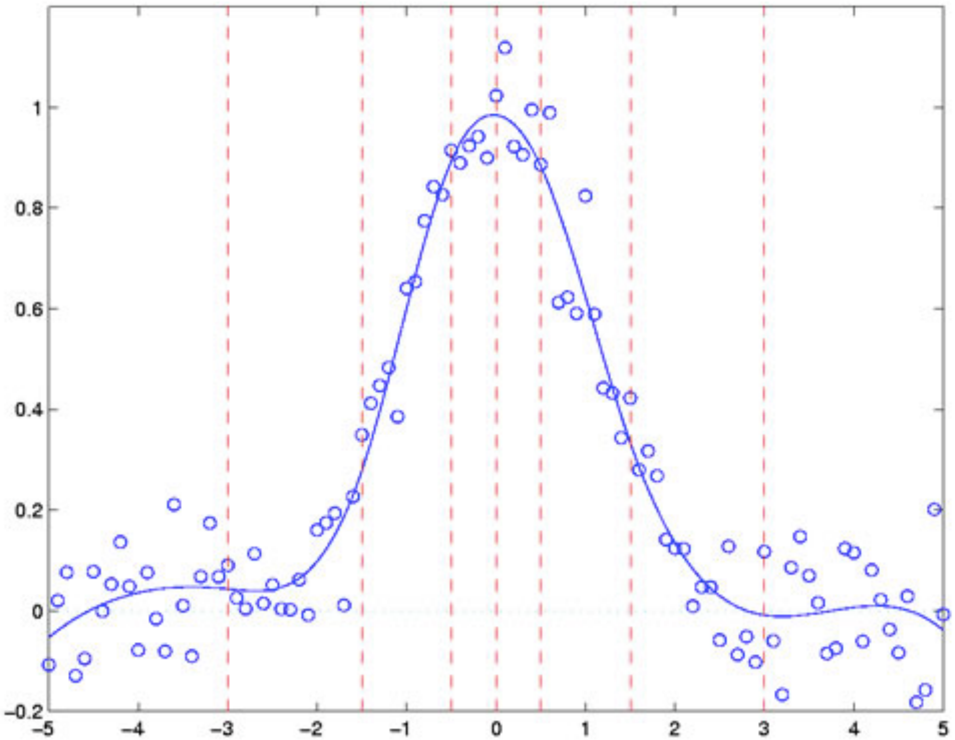




Smoothing

Spline Regression

Grace Wahba and others



Smoothing

Polynomial smoothers

Loess (Cleveland)

originally called LOWESS (Locally Weighted Scatterplot Smoothing)

renamed Loess (a wind-blown berm)

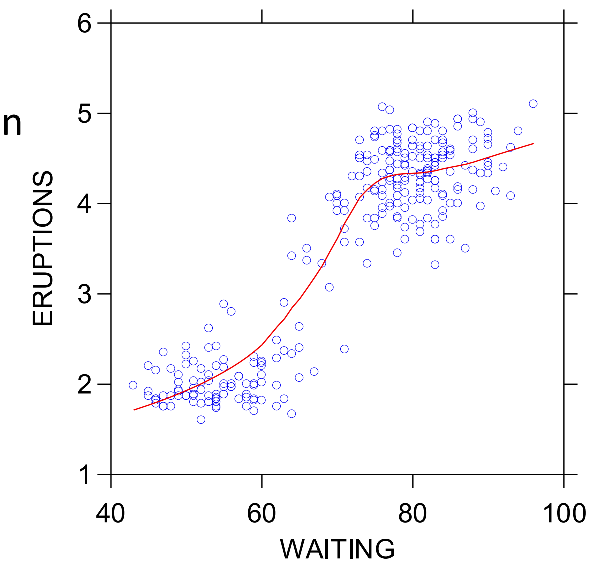
a robust hybrid of kernel and polynomial regression

Tricube kernel  *tricube*: $f(x) = a(1 - |x/w|^3)^3 : (-w \le x \le w)$, else 0

KNN window

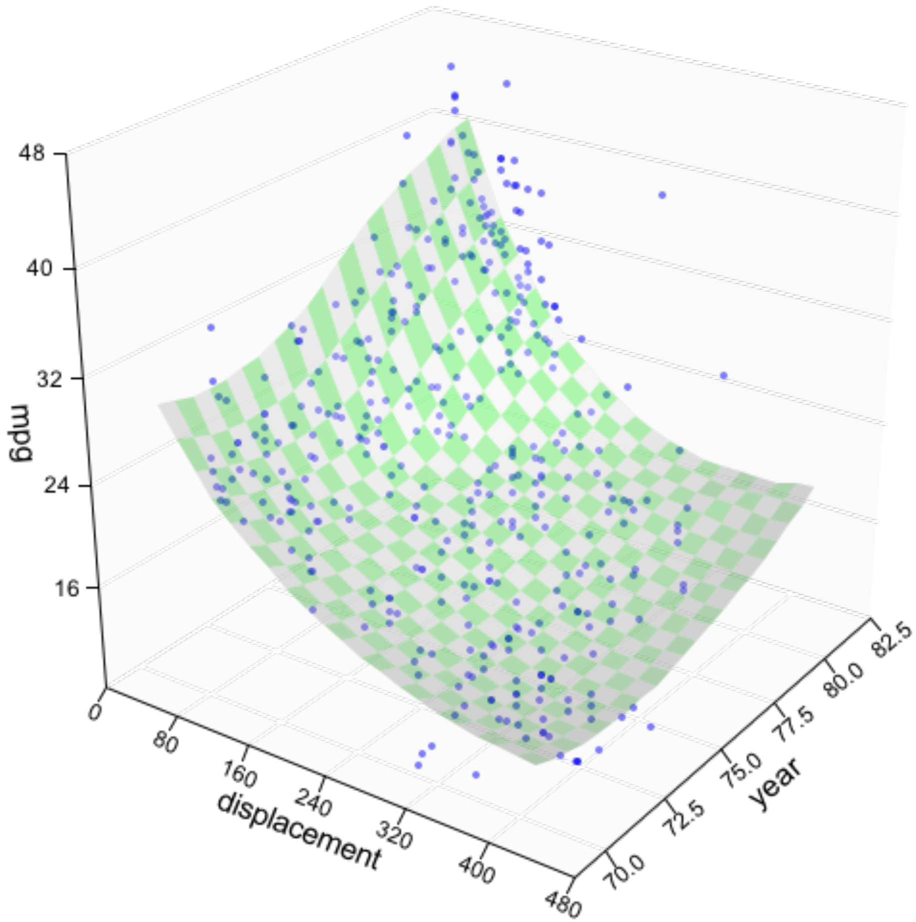
Biweight psi function

Robust linear (quadratic) regression



Smoothing

Loess



Smoothing

Principal Curves

Hastie and Stuetzle, 1989.

Each point on the curve is the average of points in a window projected onto the space of the curve.

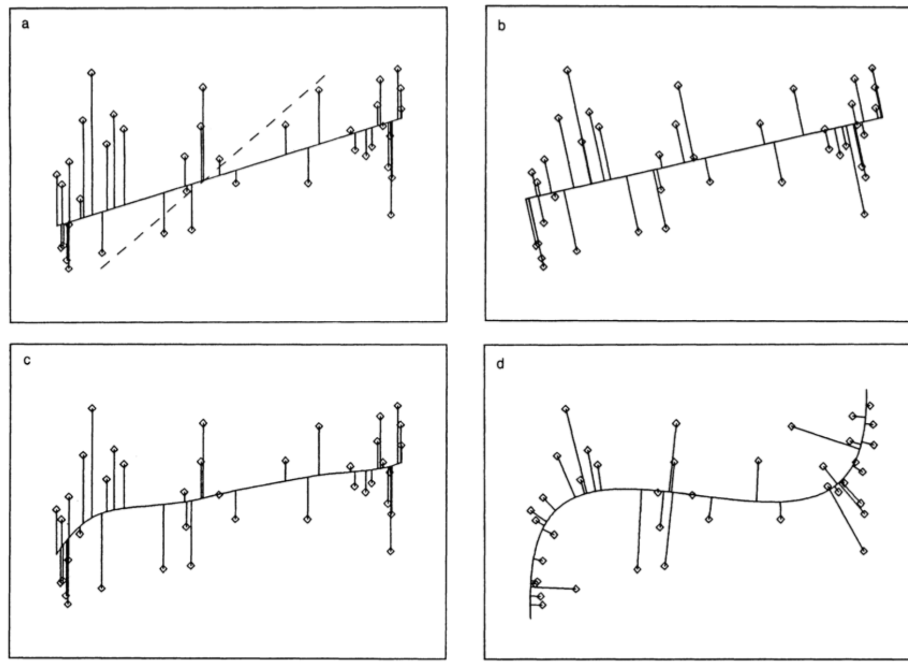


Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.

Smoothing

Smoothing Tables

Tukey Median Polish

Model: *y = row + column + rough*

1. Compute the median of each row and record the value to the side of the row. Subtract the row median from each point in that particular row.

2. Compute the median of the row medians, and record the value as the grand effect. Subtract this grand effect from each of the row medians.

3. Compute the median of each column and record the value beneath the column. Subtract the column median from each point in that particular column.

4. Compute the median of the column medians, and add the value to the current grand effect. Subtract this addition to the grand effect from each of the column medians.

5. Repeat steps 1-4 until no major changes occur with the row or column medians.

A Tukey method resurrected for RMA (Robust Microarray Average) analysis of microarrays.

Smoothing

Smoothing Tables

Tukey Median Polish

Original table

Percentage of married women by country and age

	X20.24	X25.29	X30.34	X35.39	X40.44	X45.49	X50.54
Argentina	7.5	21.1	37.1	48.4	54.9	57.9	58.4
China	32.2	77.4	92.6	95.4	95.7	94.6	92.5
Iceland	3.7	20.8	39.9	51.6	57.6	60.5	63.4
Japan	9.5	37.1	60.8	69.8	73.2	76.4	79.0
Mexico	22.2	41.4	53.4	58.8	60.4	61.3	60.0
Norway	6.3	25.0	43.7	52.0	54.8	56.7	60.7

Jim Albert http://bayes.bgsu.edu https://exploredata.wordpress.com

Smoothing

Smoothing Tables

Tukey Median Polish

Polished table

	X20.24	X25.29	X30.34	X35.39	X40.44	X45.49	X50.54
Argentina	2.6	21.8	40.8	49.4	52.8	55.1	58.3
China	45.6	64.8	83.8	92.4	95.8	98.1	101.3
Iceland	4.8	23.9	43.0	51.6	54.9	57.3	60.4
Japan	23.0	42.1	61.1	69.8	73.1	75.5	78.6
Mexico	12.0	31.1	50.1	58.8	62.1	64.5	67.6
Norway	5.2	24.3	43.4	52.0	55.3	57.7	60.8

Smoothing

Smoothing Tables

Tukey Median Polish

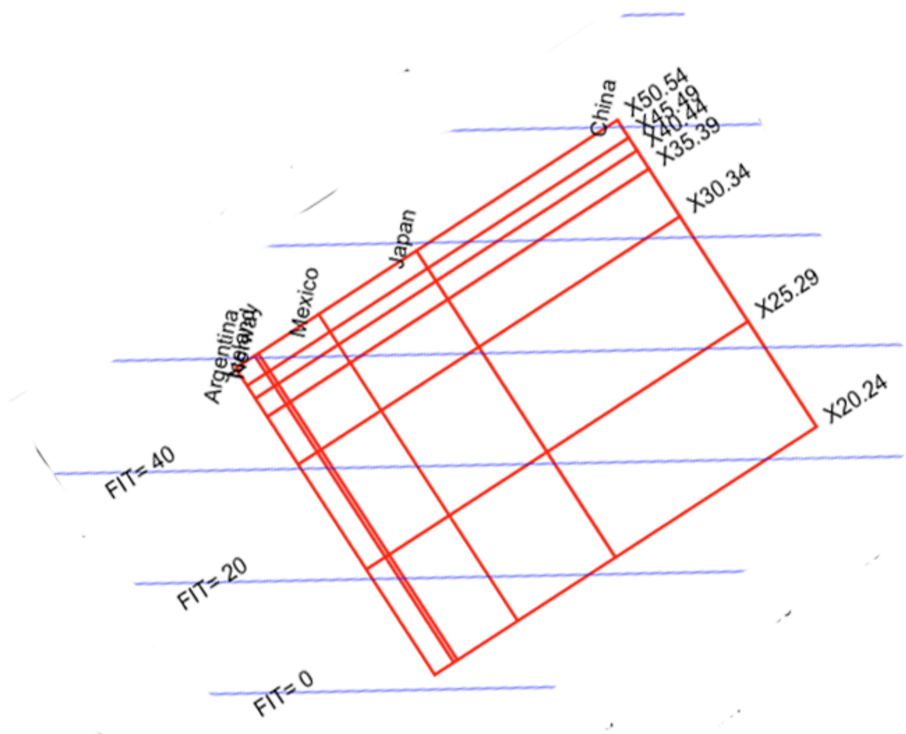
Residuals

	X20.24	X25.29	X30.34	X35.39	X40.44	X45.49	X50.54
Argentina	4.9	-0.7	-3.7	-1	2.1	2.8	0.1
China	-13.4	12.6	8.8	3	-0.1	-3.5	-8.8
Iceland	-1.1	-3.1	-3.1	0	2.7	3.2	3.0
Japan	-13.5	-5.0	-0.4	0	0.1	1.0	0.4
Mexico	10.2	10.3	3.2	0	-1.7	-3.2	-7.6
Norway	1.1	0.7	0.4	0	-0.5	-1.0	-0.1

Smoothing

Smoothing Tables

Tukey Median Polish



Jim Albert -- https://exploredata.wordpress.com

Smoothing

Smoothing Tables

Conjoint Measurement

Luce, R.D., Tukey, J.W. (1964). Simultaneous conjoint measurement: A new scale type of fundamental measurement. *Journal of Mathematical Psychology 1,* 1–27.

This paper refuted the longstanding claim among physicists that *fundamental measurement* (*concatenation of measures*) was the only admissible measurement foundation for science.

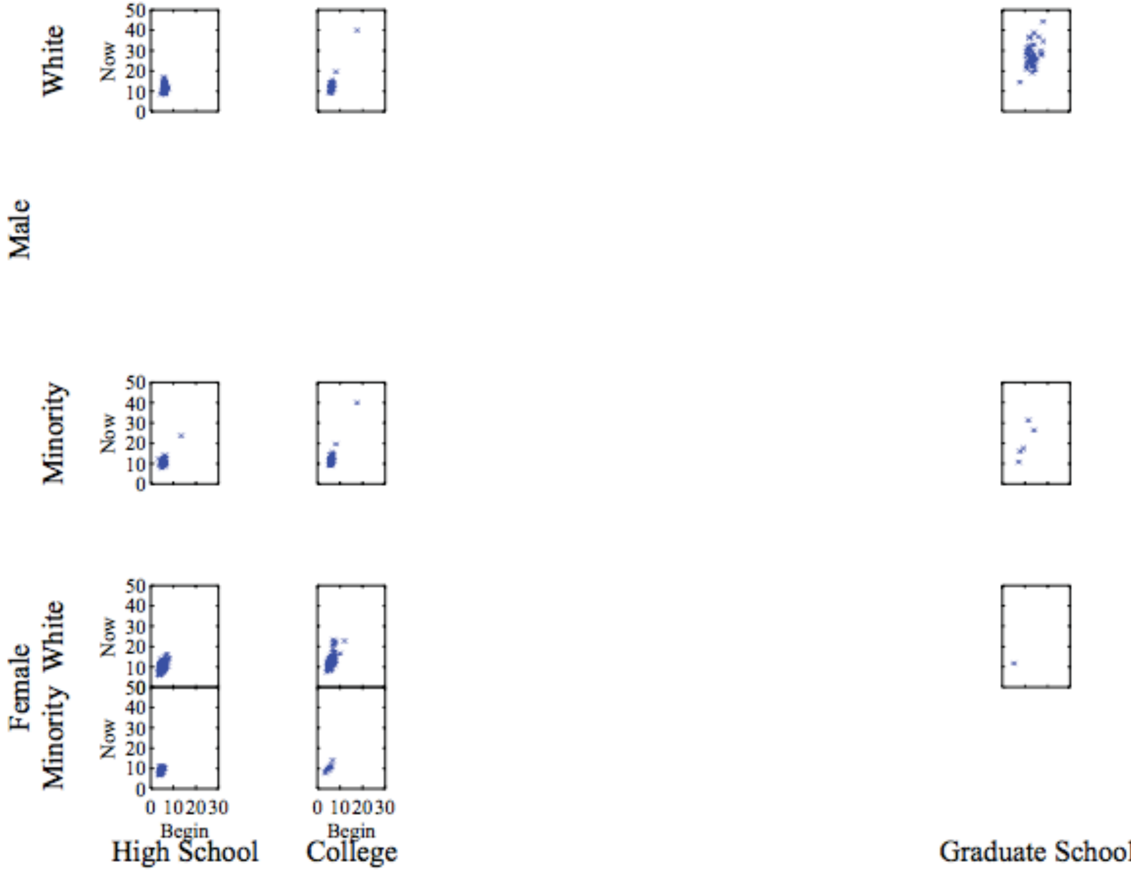
Tukey's median polish is an implementation of conjoint measurement.

Market researchers adopted the idea and called it *conjoint analysis*, but they threw out the baby with the bathwater

They used an ordinary analysis of variance model instead of nonmetric techniques

Smoothing

Conjoint Measurement



Smoothing

Smoothing Tables

- Head Injury Index
- Frontal crashes
- Cars and Trucks
- NHTSA

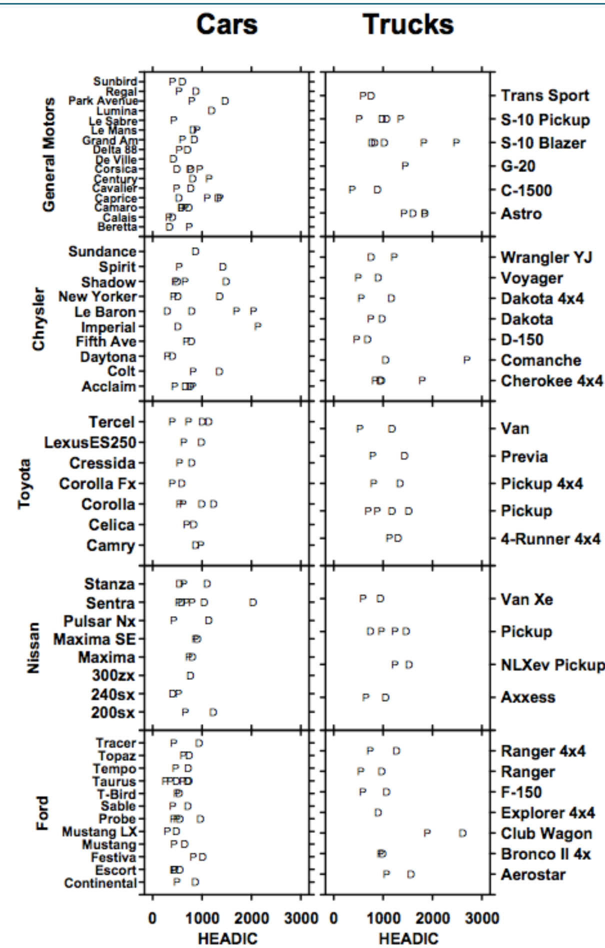


Figure 16.2 *Estimated crash head injury criterion (P=passenger, D= driver)*

Smoothing

Smoothing Tables

$$H = C + M + V + O + T(MV) + MV + MO + VO + OT(MV) + MVO$$

H : Head Injury Index
C : constant term (grand mean)
M : Manufacturer
V : Vehicle (car/truck)
O : Occupant (driver/passenger)
T : Model

Be careful if you drive a truck

- Body-on-frame rigid frame rails dangerous
- Don't absorb shock in head-on collision

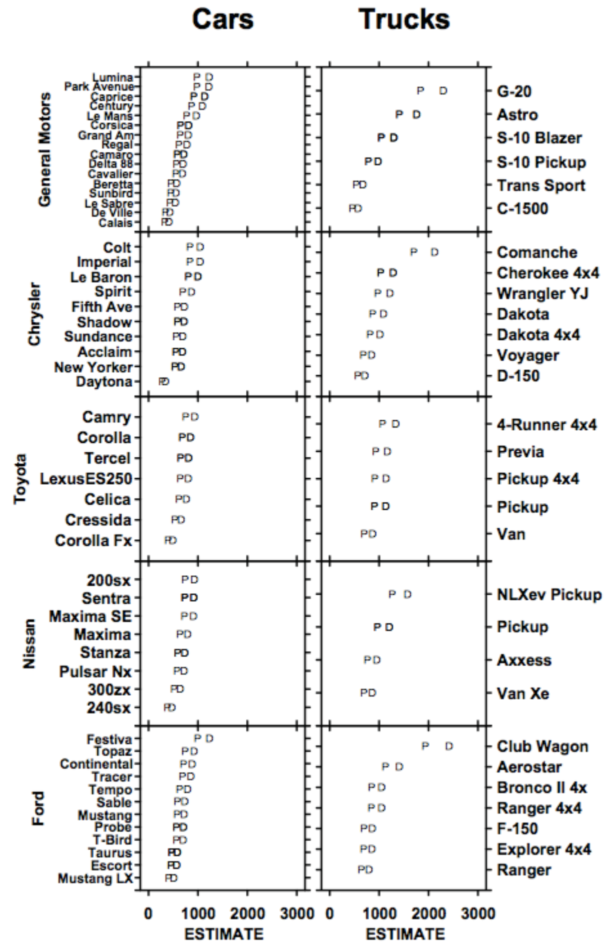


Figure 16.3 *Subset model for crash data sorted by estimate*

Smoothing

References

Andrews, D., P. Bickel, F. Hampel, P. Huber, W. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations, *Journal of the Royal Statistical Society*, pp. 211-243, discussion pp. 244-252.

Hastie, T., and Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association, 84*, 502-516.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics, 2nd ed.*, Wiley.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression*. Addison-Wesley.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Velleman, P. and Hoaglin, D. (1981). *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.