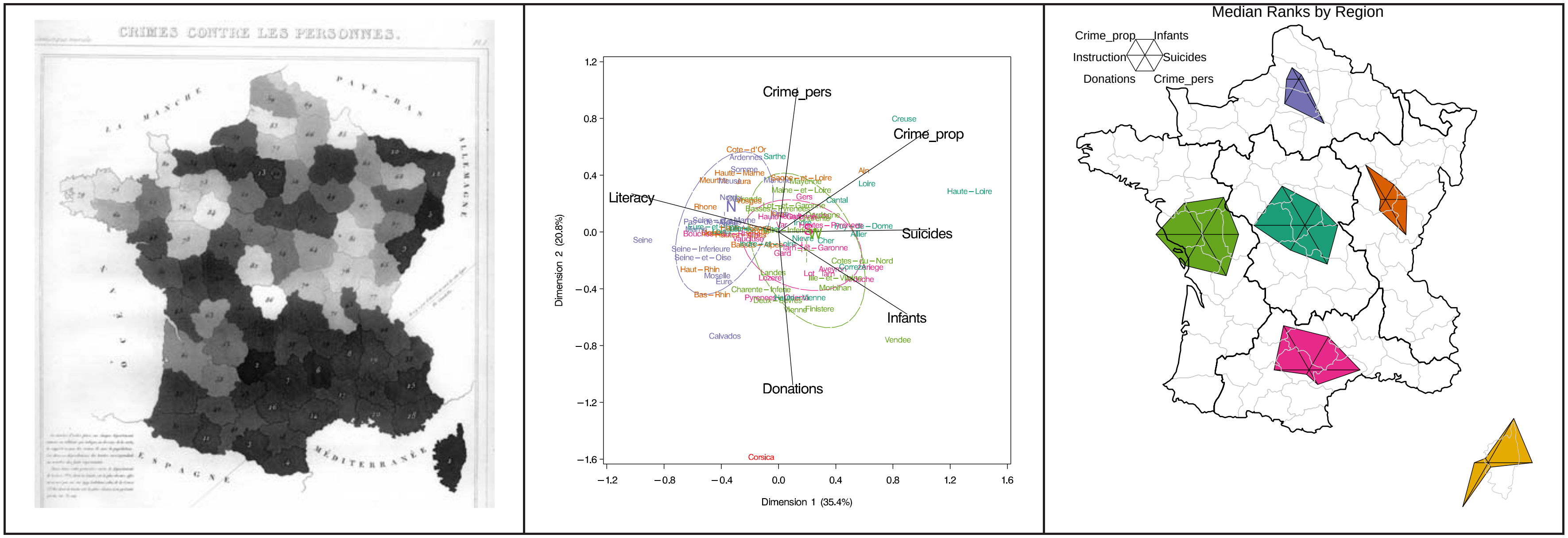
Visualizing Multivariate Uncertainty:

Some Graphical Methods for Multivariable Spatial Data



Michael Friendly

York University

`http://www.math.yorku.ca/SCS/friendly.html`

National Academy of Sciences

Washington, DC, Mar, 2005

Outline

- **Multivariate uncertainty and “moral statistics”**
  - A. M. Guerry’s *Moral Statistics of France*
  - Guerry’s data and analyses
- **Multivariate analyses**: Data-centric displays
  - Bivariate plots and data ellipses
  - Biplots
  - Canonical discriminant plots
  - HE plots for multivariate linear models
- **Multivariate mapping**: Map-centric displays
  - Star maps
  - Reduced-rank color maps

Multivariate Uncertainty and “Moral Statistics” $\sim$ 1800

It is a capital mistake to theorize before one has data. Sherlock Homes in *Scandal in Bohemia*

- **What to do about crime?**
  - Liberal view: increase education, literacy
  - Conservative view: build more prisons
- **What to do about poverty?**
  - Liberal view: increase social assistance
  - Conservative view: build more poor-houses
- **But**:
  - Little actual data – all armchair theorizing
  - No ways to understand or visualize *relationships* between variables
    - Statistical graphics just invented (Playfair)— line graph, bar chart, pie chart
    - All 1D or 1.5D (time series)

The rise of “moral statistics” and modern social science

- **Political arithmetic**: William Petty (and others)
  - 1654— first attempt at scientific survey (on Irish estates)
  - 1687— idea that wealth and strength of a state depended on its subjects (number and characteristics)
- **Demography**: Johann Peter Süssmilch (1741)—
  - importance of measuring and analyzing population distributions
  - idea that ethical and state policies could encourage growth and wealth (increase birth rate, decrease death rate)
    - discourage alcohol, gambling, prostitution & priestly celibacy
    - encourage state support for medical care, distribution of land, lower taxes
- **Statistik**: Numbers of the state (1800–1820), Germany and France
  - collect data on imports, exports, transportation, ...
- **Guerry & Quetelet**
  - Quetelet: Concepts of “average man” and “social physics”
  - Guerry: First real social data analysis (Guerry, 1833)

Guerry's data

- **Compte général** de l'administration de la justice criminelle en France
  - The first national compilation of official justice data (1825)
    - detailed data on all charges and disposition
    - collected quarterly in all 86 departments.
  - Other sources: Bureau de Longitudes (illegitimate births); Parent-Duchâtelet (prostitutes in Paris); Compte du ministere du guerre (military desertions); ...
- **Moral variables**: Scaled so 'more' is 'better'

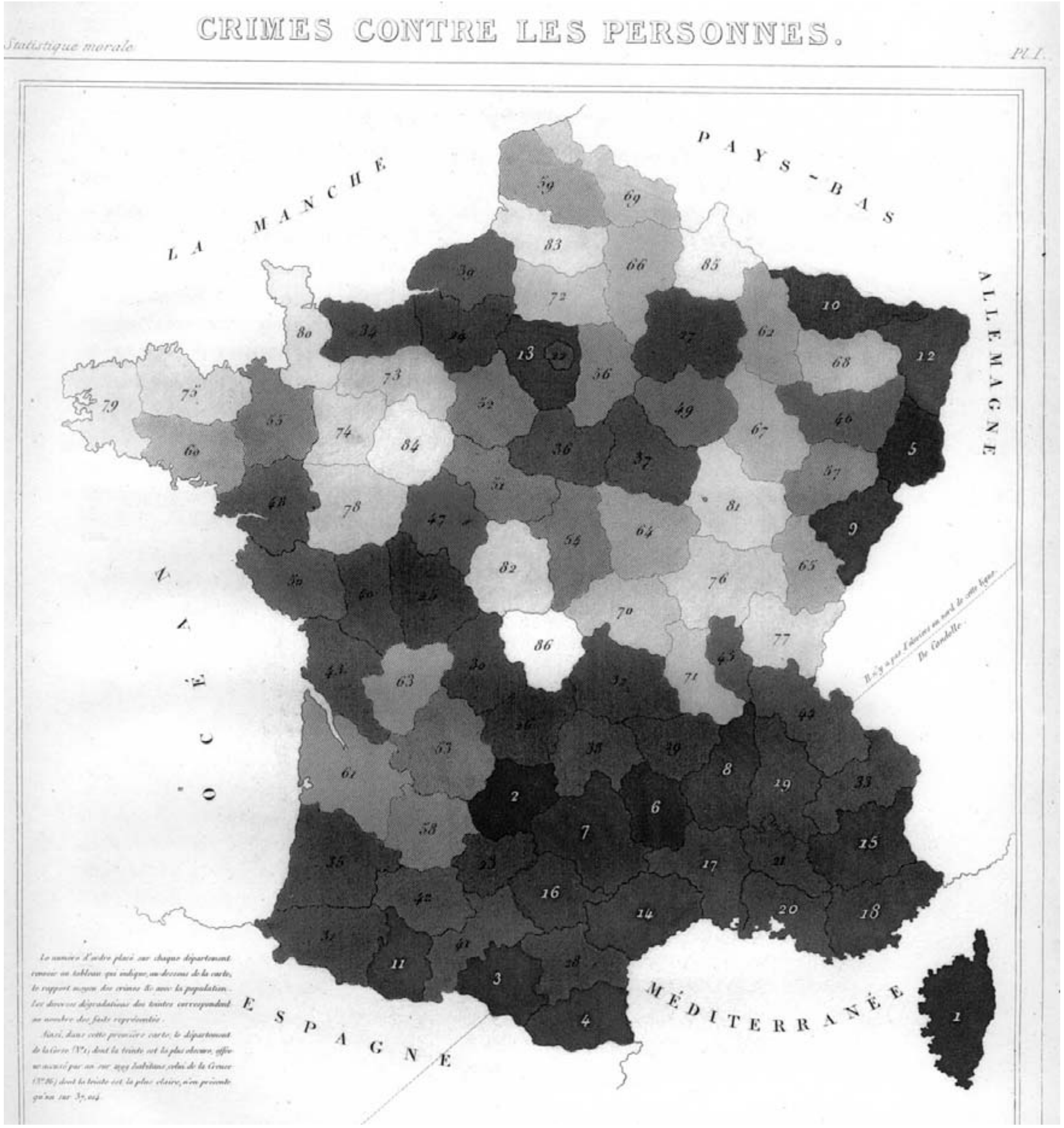
`Crime_pers`	Population per Crime against persons
`Crime_prop`	Population per Crime against property
`Donations`	Donations to the poor
`Infants`	Population per illegitimate birth
`Literacy`	Percent who can read & write
`Suicides`	Population per suicide

  - Tried to define these to ensure comparability and representativeness
    - Crime: Use number of *accused* rather than *convicted*
    - Literacy: Reported levels of education unreliable; use data from military draft examinations (% of young men able to read and write)
- **Other variables**: Ranks by department: wealth, commerce, ...

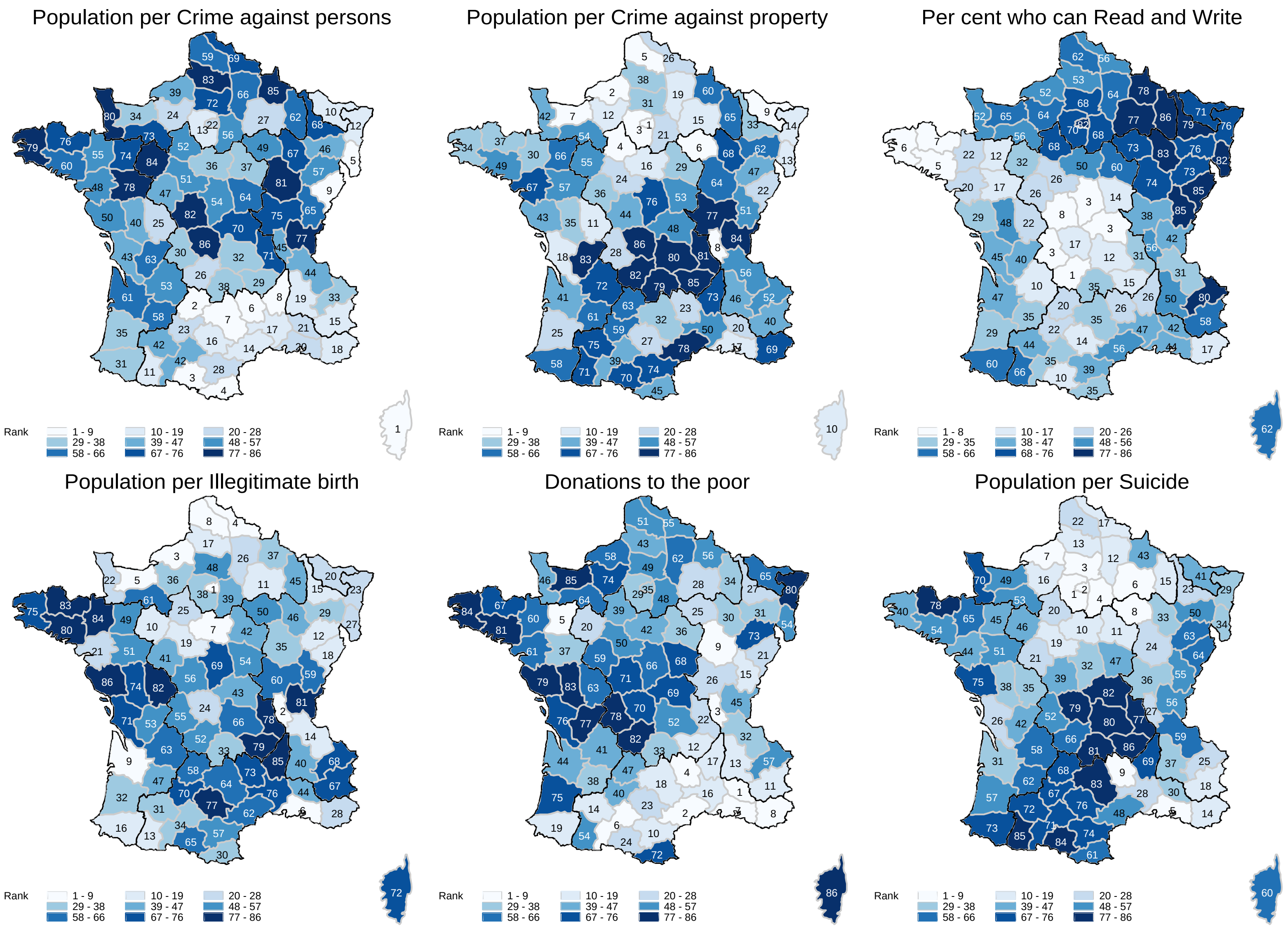
Guerry's Questions

- Should crime and other moral variables be considered as structural, lawful characteristics of society, or simply as indicants of individual behavior?
  - Statistical regularity as the key to social science ("social physics") social equivalent of "law of large numbers")
  - Guerry showed that rates of crime had nearly invariant distributions over time (1825–1830) when classified by region, sex of accused, type of crime, etc. *"We would be forced to recognze that the facts of moral order, like those of physical order, obey invariant laws..."* (p.14)
- Relations between crime and other moral variables
  - Do crimes against persons and crimes against property show the same or different trends?
  - How does crime relate to education and literacy?
    - Some "armchair" arguments had suggested increasing literacy to decrease crime: *"The definitive result shows that 67 out of 100 prisoners can neither read nor write. What stronger proof could there be that ignorance is the mother of all vices"* (A. Taillander, 1828)
  - Does crime vary coherently over regions of France (C, N, S, E, W)?

Guerry’s maps


CRIMES CONTRE LES PERSONNES.
Statistique morale
Pl. I.
LA MANCHE
PAYS-BAS
ALLEMAGNE
OCÉAN
ESPAGNE
MÉDITERRANÉE
Il n'y a pas d'oliviers au nord de cette ligne. De Candolle.
59 69 83 66 85 39 72 10 80 34 24 13 22 56 27 62 68 12
79 75 55 73 52 49 46 60 74 84 36 37 67 5 48 51 81 57
78 47 64 9 82 54 76 65 50 40 25 70 86 77 44 63 30 32 71 43
61 53 26 38 29 8 19 33 2 6 7 15 58 35 23 17 21 42 16 14 20 18
31 11 41 3 28 4 1
Le numéro d'ordre placé sur chaque département renvoie au tableau qui indique, au-dessous de la carte, le rapport moyen des crimes &c avec la population. Les diverses dégradations des teintes correspondent au nombre des faits représentés.
Ainsi, dans cette première carte, le département de la Corse (N°1), dont la teinte est la plus obscure, offre un accusé par un sur 2199 habitans, celui de la Creuse (N°86) dont la teinte est la plus claire, n'en présente qu'un sur 37,014.

Guerry’s maps



Population per Crime against persons

Population per Crime against property

Per cent who can Read and Write

Population per Illegitimate birth

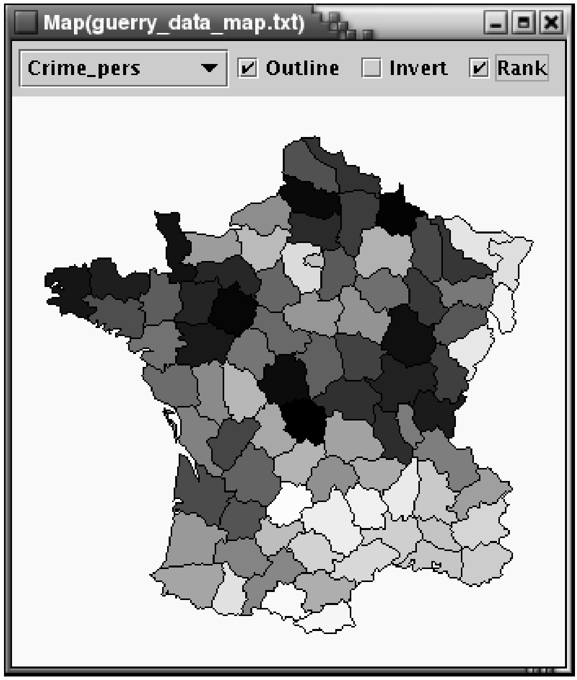
Donations to the poor

Population per Suicide

Guerry's analyses

Relate variables by comparing maps and ranked lists (1st $||$ coordinate plot)

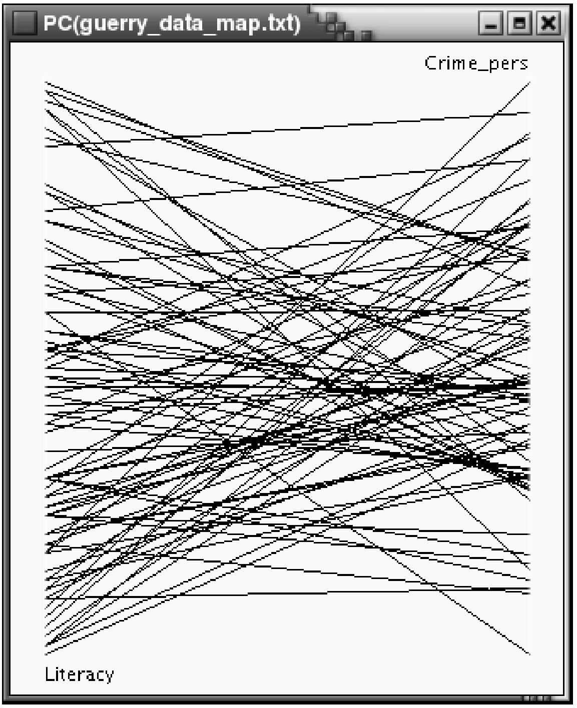
- Conclusion: no clear relation between crime and literacy



Map(guerry_data_map.txt)

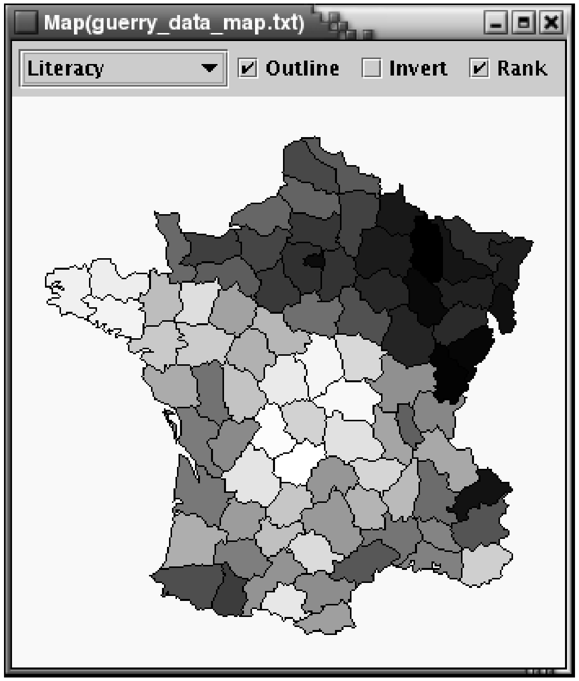
Crime_pers ☑ Outline □ Invert ☑ Rank

Literacy



PC(guerry_data_map.txt)

Ranked lists



Map(guerry_data_map.txt)

Literacy ☑ Outline □ Invert ☑ Rank

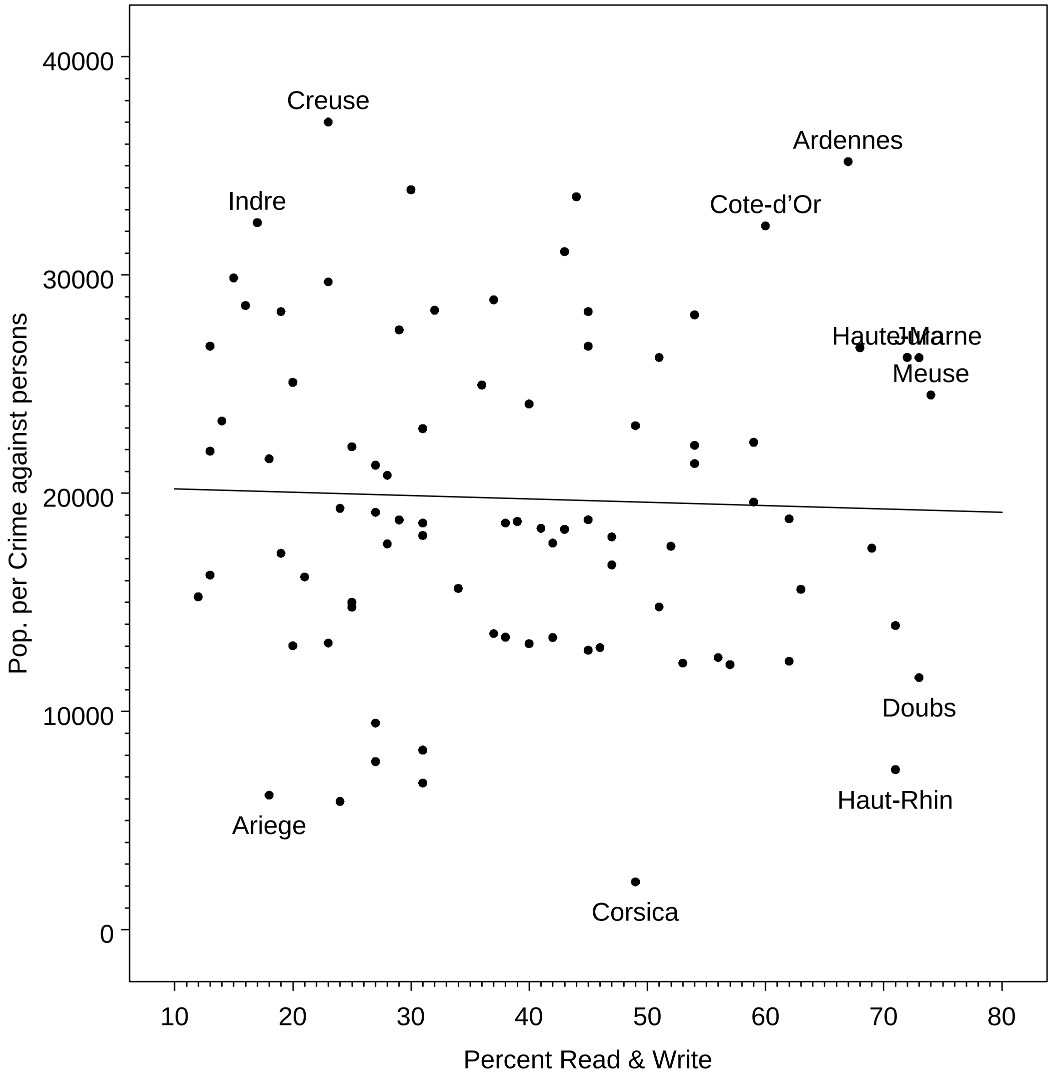
Crimes against persons

- Similar analyses for other variables (suicide, illegitimate births, ...)

Graphical methods for multivariate data

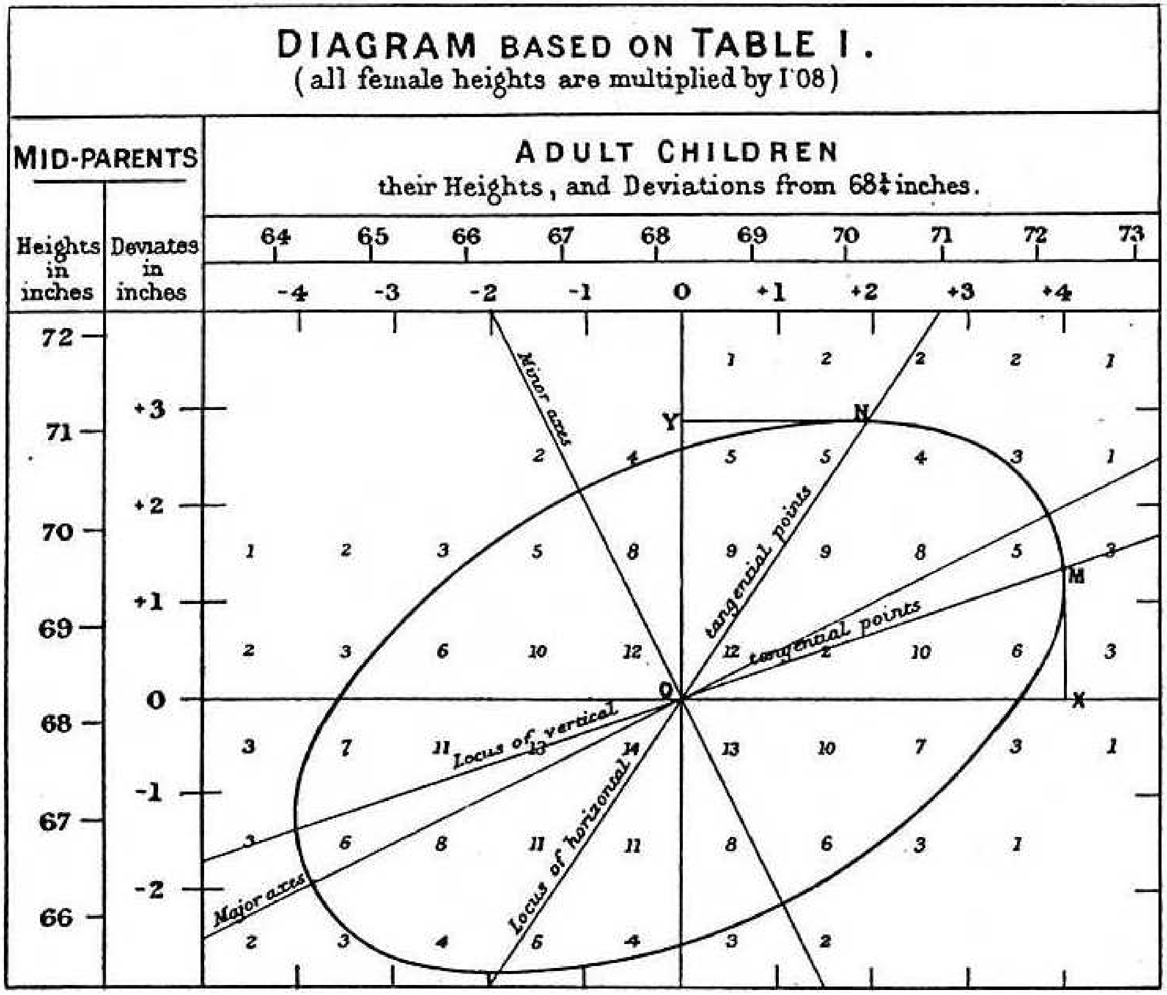
- **Bivariate displays**: Bivariate displays can be enhanced to show statistical relations more clearly and effectively
  - Scatterplots with data (concentration) ellipses and smoothed (loess) curves
  - Scatterplot matrices
  - Corrgrams and visual thinning
- **Reduced-rank displays**: Multivariate visualization techniques can show the statistical data in simple ways, using dimension reduction techniques.
  - Biplots - show variables and observations in space accounting for greatest variance
  - Canonical discriminant plots - show variables and observations in space accounting for greatest between-group variation
- **HE plots**: Visualization for Multivariate Linear Models

Bivariate plots: Points and visual summaries



Scatterplot with linear regression line

The Data Ellipse: Galton's Discovery



Pearson (1920): "... one of the most noteworthy scientific discoveries arising from pure analysis of observations."

The Data Ellipse: Details

Visual summary for bivariate marginal relations

- **Shows**: means, standard deviations, correlation, regression line(s)
- **Defined**: set of points whose squared Mahalanobis distance $\leq c^2$,

$$D^2(\boldsymbol{y}) \equiv (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T}\, \boldsymbol{S}^{-1}\, (\boldsymbol{y} - \bar{\boldsymbol{y}}) \leq c^2$$

  $\boldsymbol{S} =$ sample variance-covariance matrix

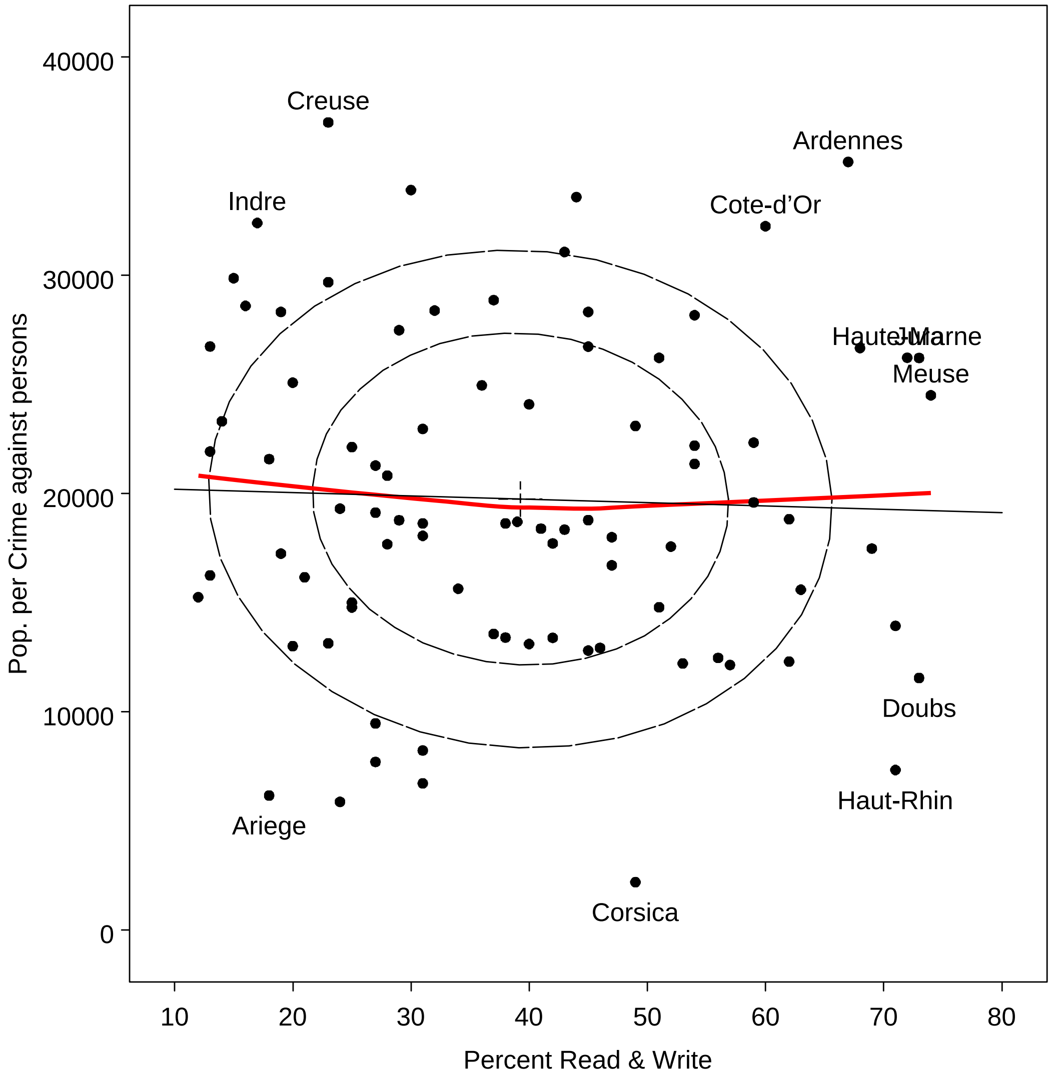
- **Radius**: when $\boldsymbol{y}$ is approx. bivariate normal, $D^2(\boldsymbol{y})$ has a large-sample χ^2_2 distribution with 2 degrees of freedom.
  - $c^2 = \chi^2_2(0.40) \approx 1$: 1 std. dev univariate ellipse– 1D shadows: $\bar{y} \pm 1s$
  - $c^2 = \chi^2_2(0.68) = 2.28$: 1 std. dev bivariate ellipse
  - Small samples: $c^2 \approx 2F_{2,n-2}(1-\alpha)$
- **Construction**: Transform the unit circle, $\mathcal{U} = (\sin \boldsymbol{\theta}, \cos \boldsymbol{\theta})$,

$$\mathcal{E}_c = \bar{\boldsymbol{y}} + c\boldsymbol{S}^{1/2}\mathcal{U}$$

  $\boldsymbol{S}^{1/2} =$ any “square root” of $\boldsymbol{S}$ (e.g., Cholesky)

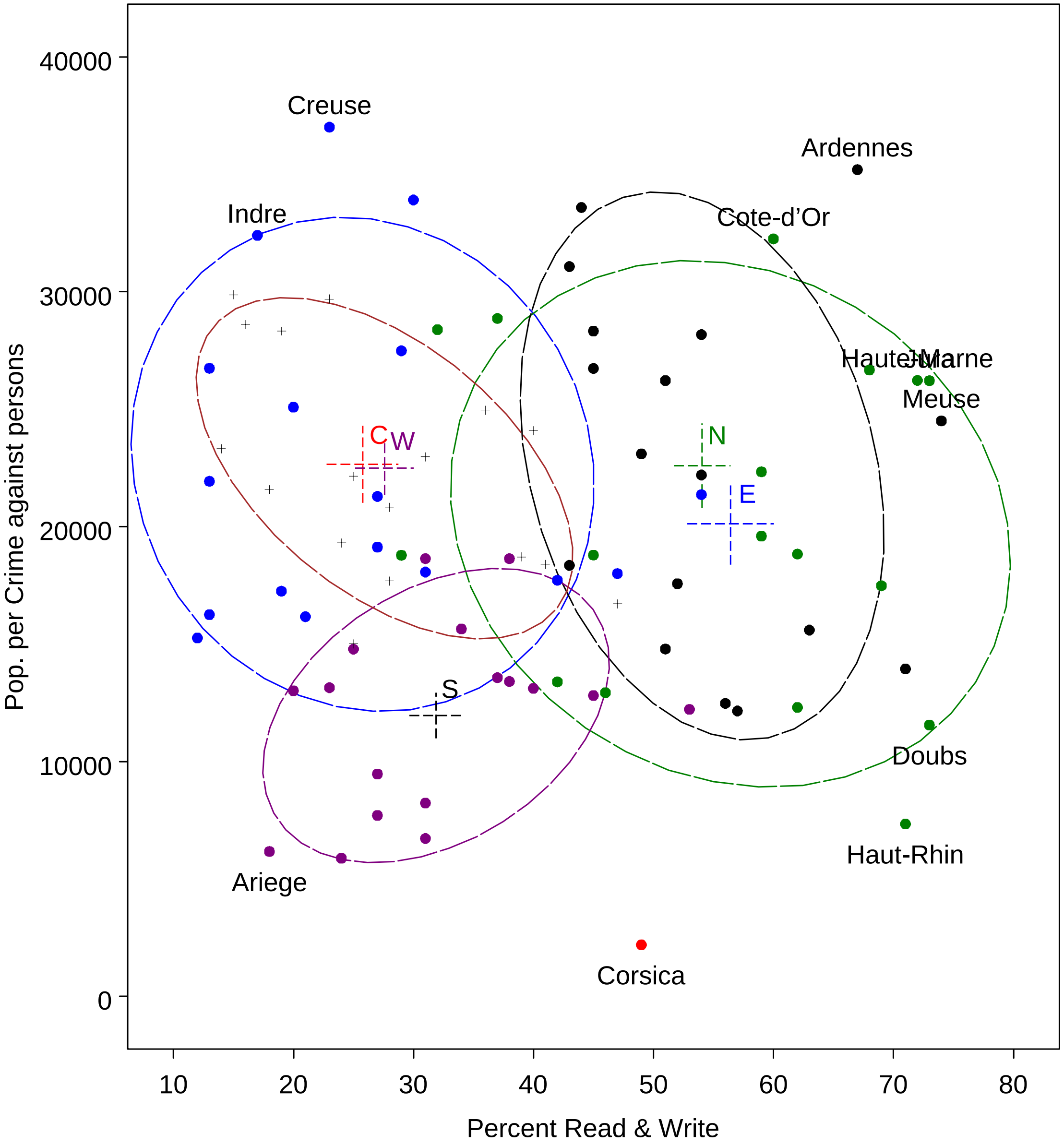
- **Robust version**: Use robust covariance estimate (MCD, MVE)
- **Nonparametric version**: Use kernel density estimation

Bivariate plots: Data ellipse and smoothing

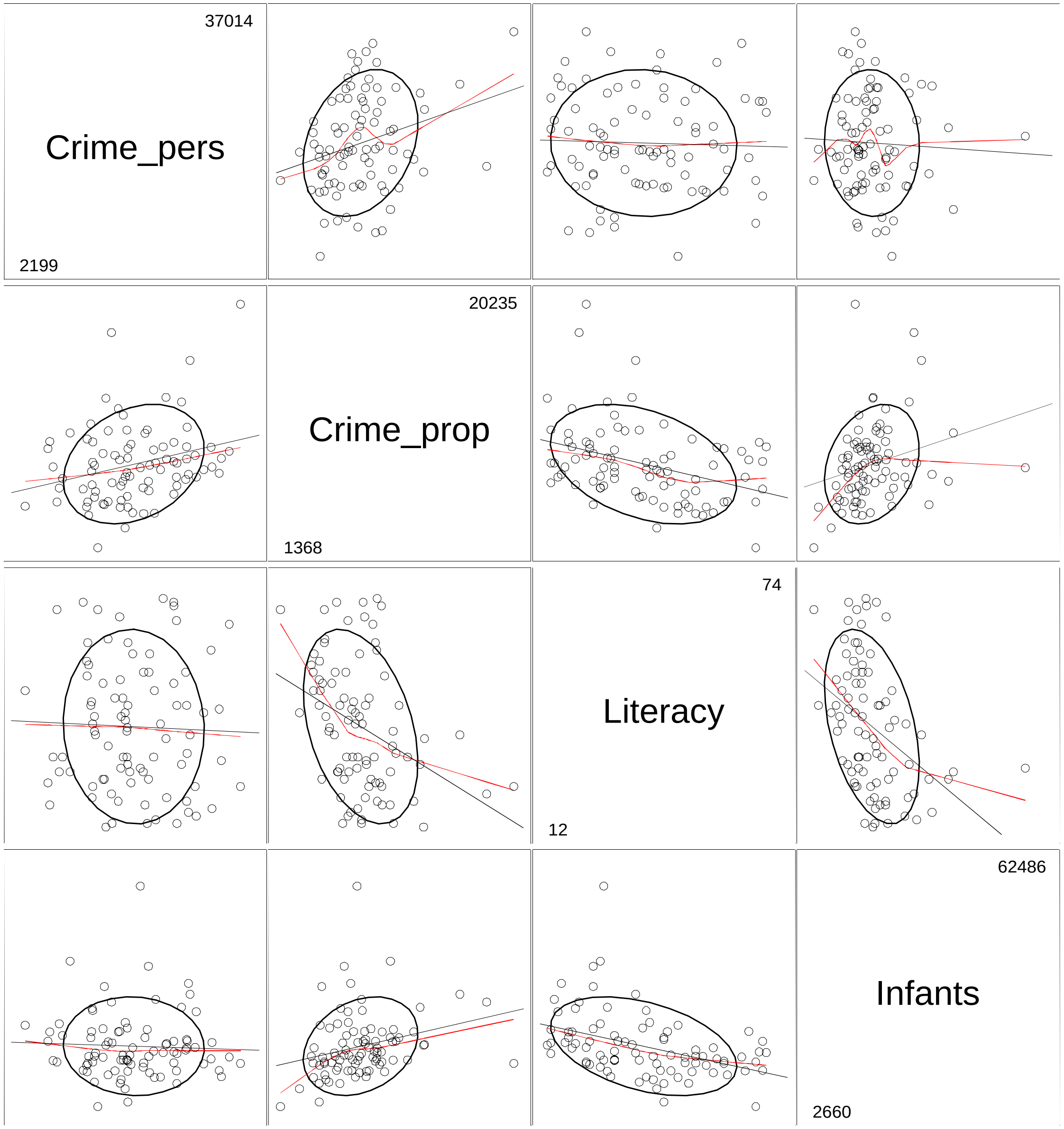


Scatterplot with 68% data ellipse and smoothed (loess) curve

Bivariate plots: Region differences

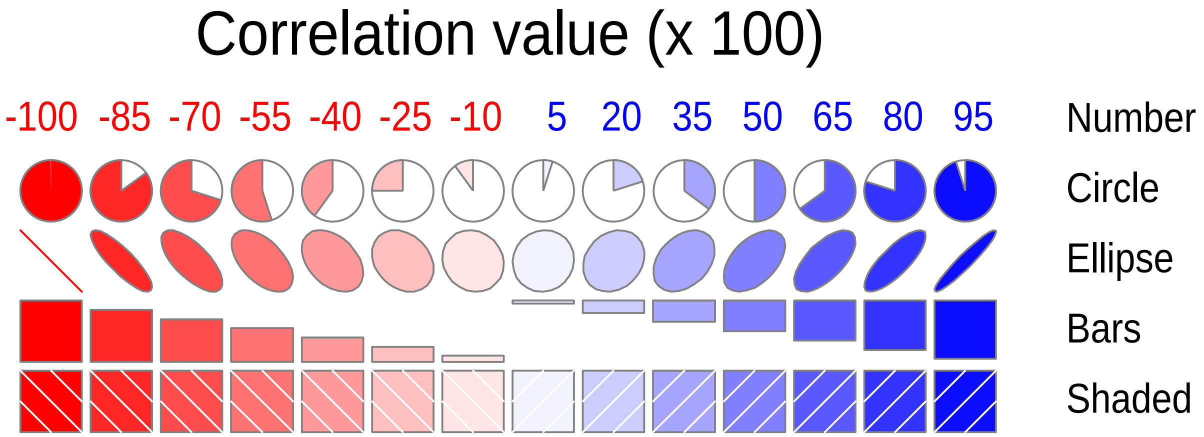


Bivariate plots: Scatterplot matrices



Corrgrams— Correlation matrix displays

- How to show a correlation matrix for different purposes? (Friendly, 2002)
- Render a correlation to depict sign and magnitude (tasks: lookup, comparison, detection)

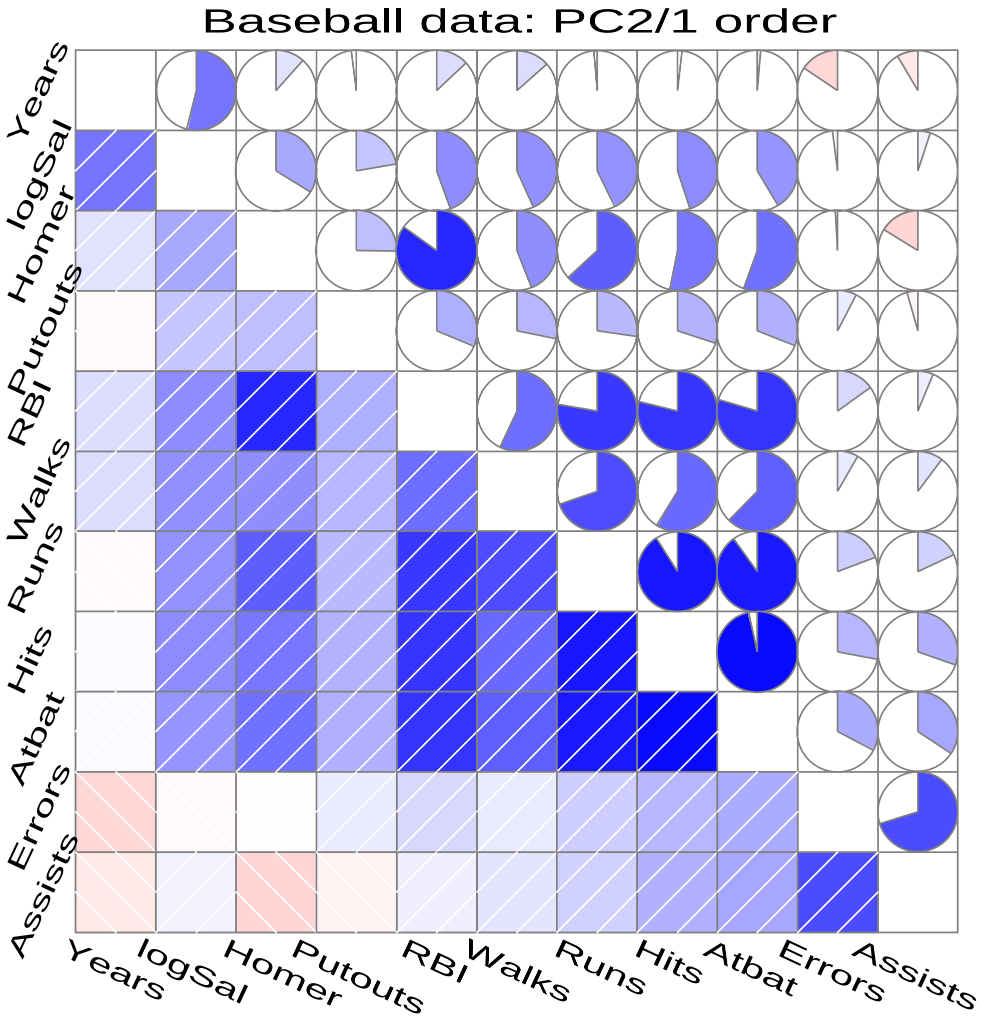


Task-specific renderings:

Task	Lookup	Comparison	Detection
Rendering	Number	Circle	Shading

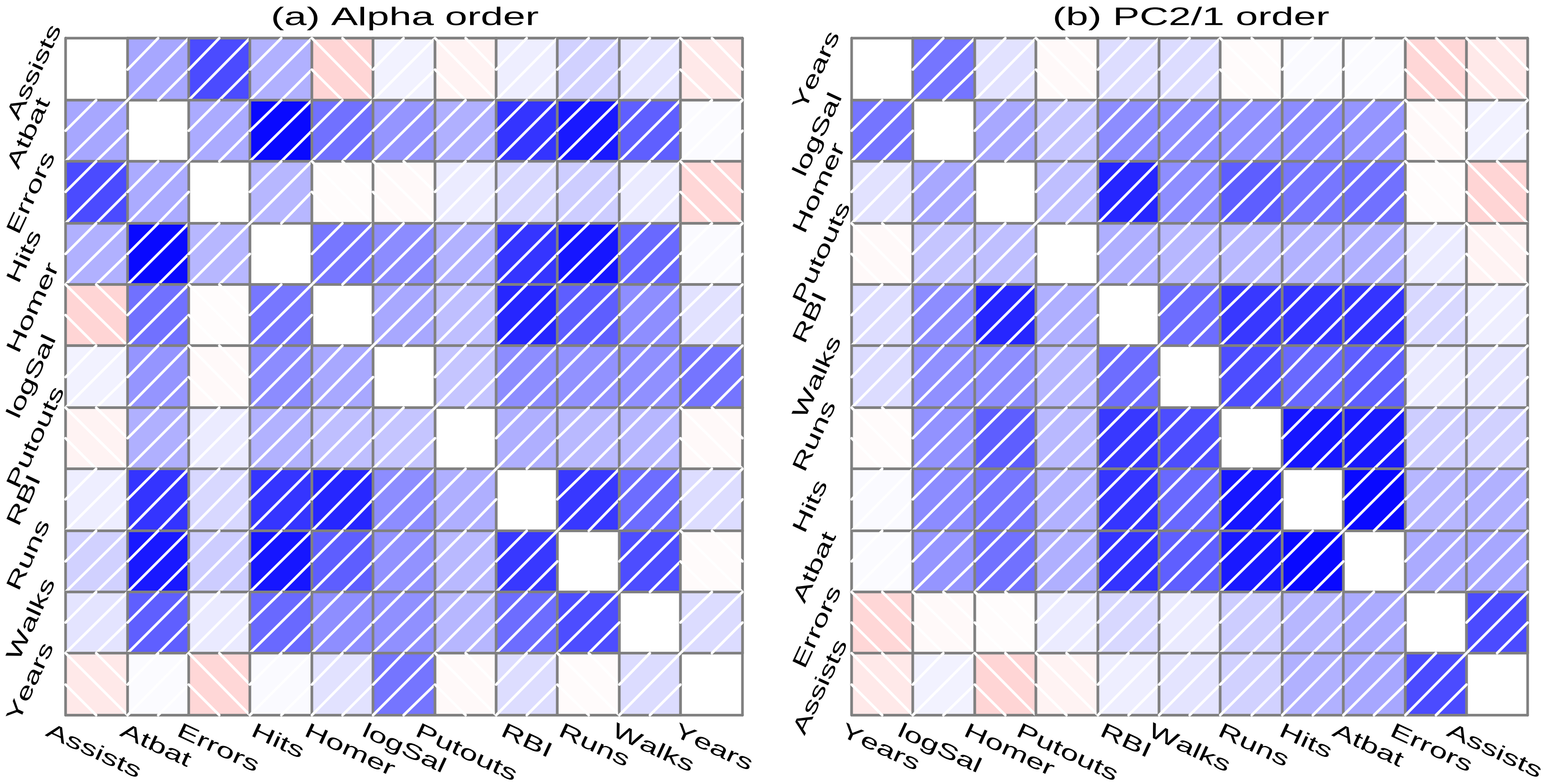
Corrgrams— Rendering

Baseball data: (lower) Patterns vs. (upper) comparison



Corrgrams— Variable ordering

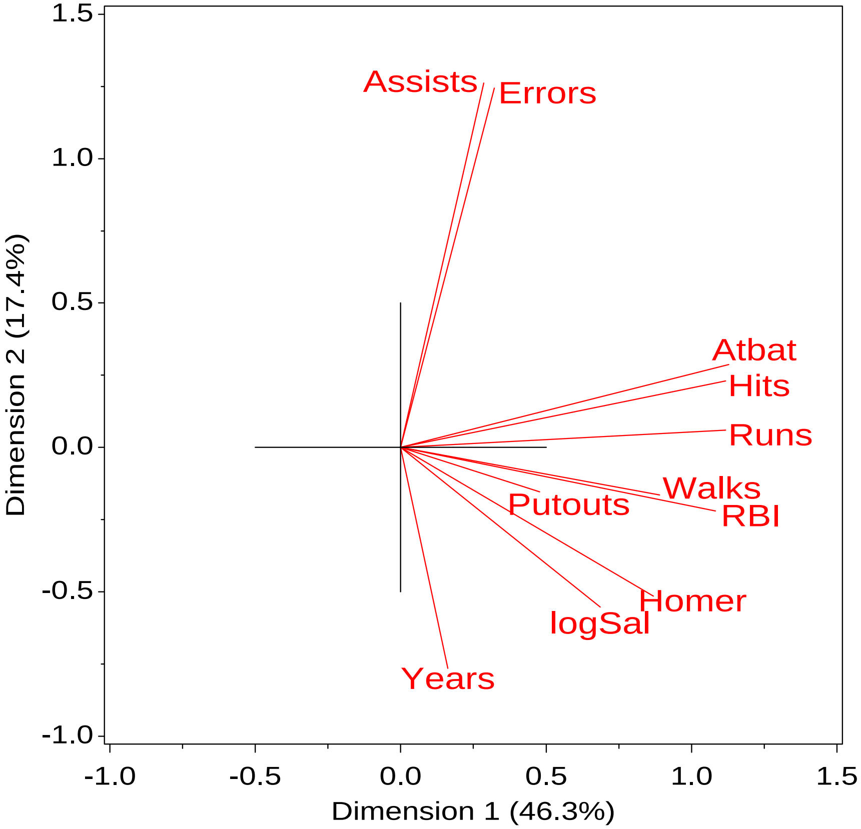
Baseball data: (a) alpha vs. (b) correlation ordering (Friendly and Kwan, 2003)



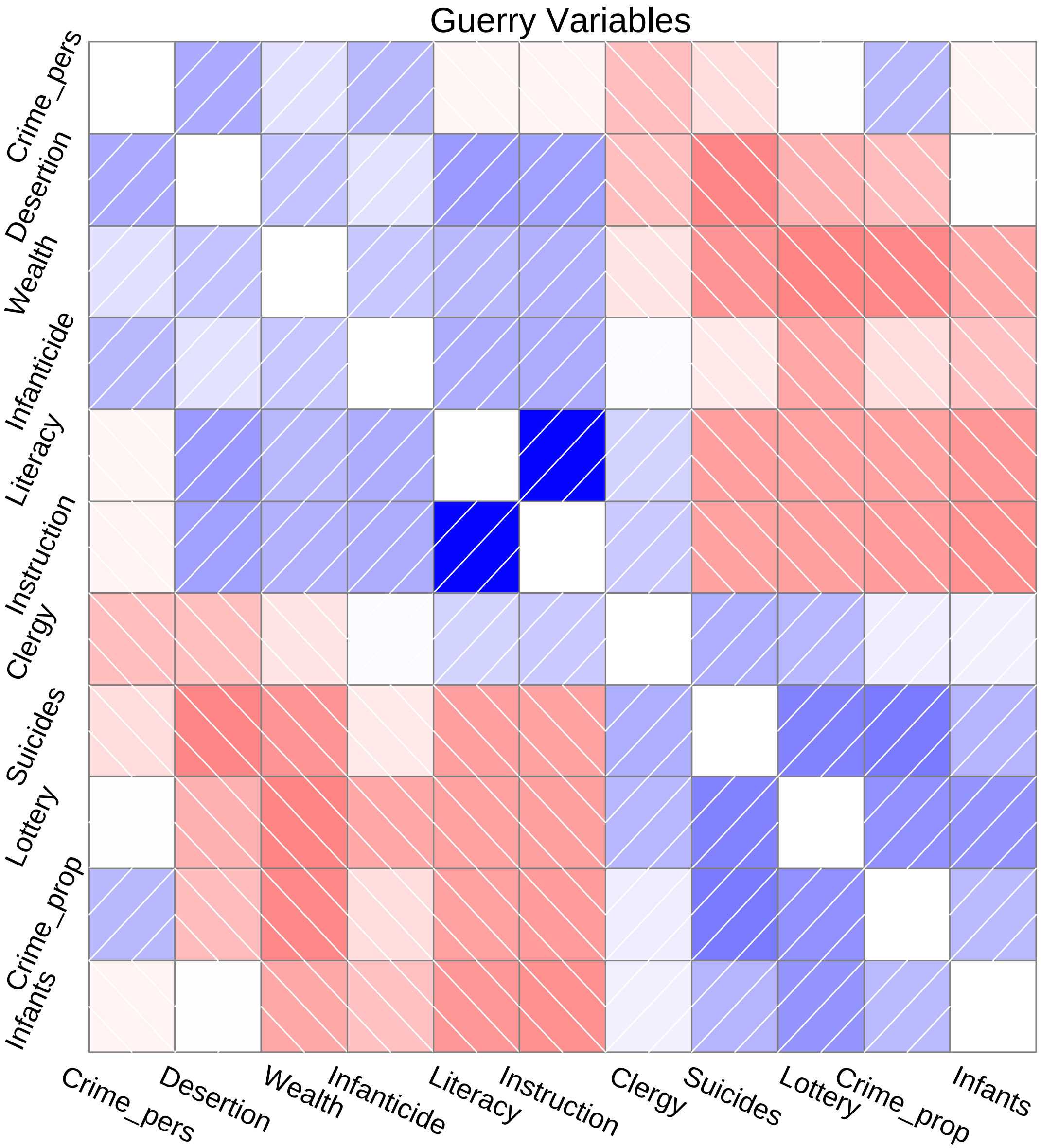
See: `http://www.math.yorku.ca/SCS/sasmac/corrgram.html`

Corrgrams— Variable ordering

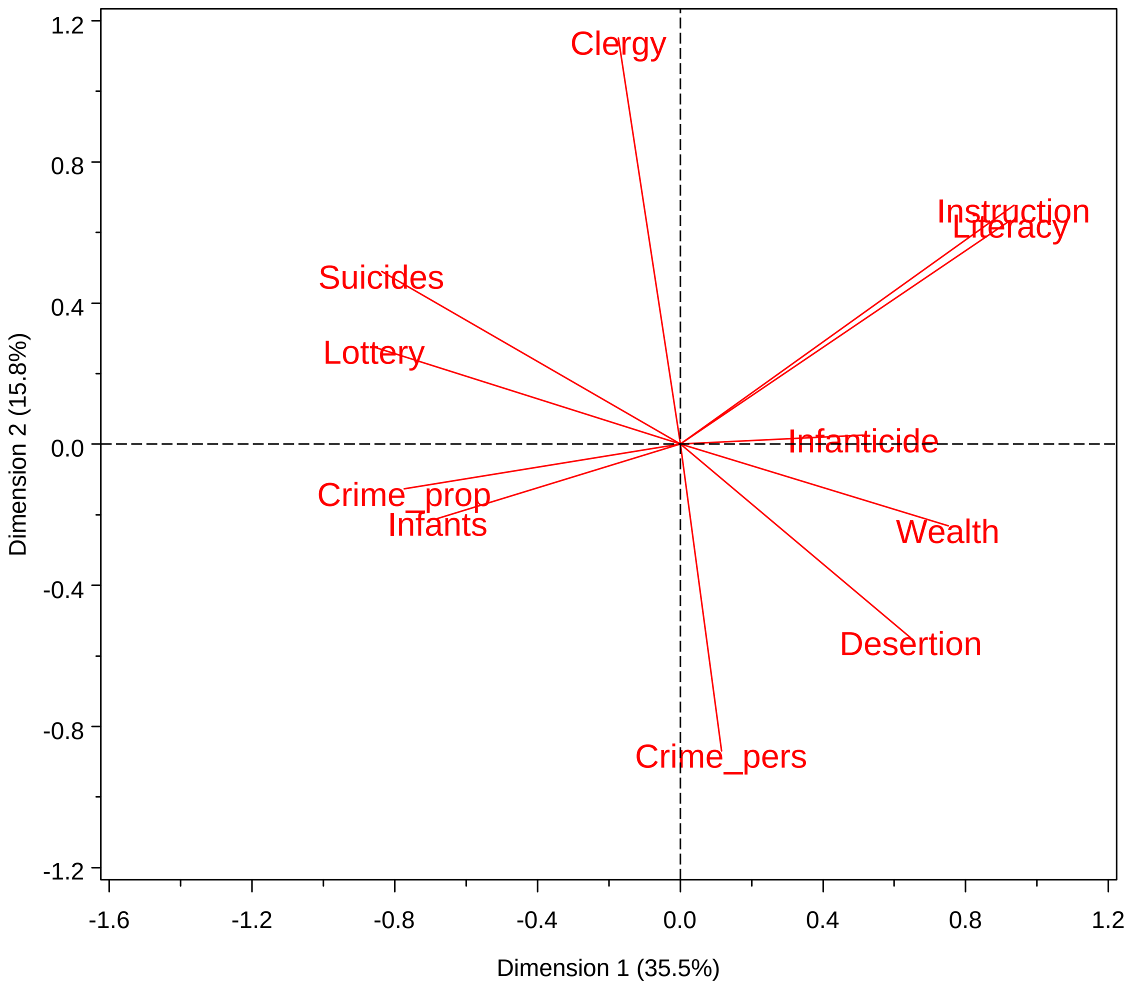
- Reorder variables to show similarities: PC1 or angles (PC2/PC1)



Corrgrams— Guerry data

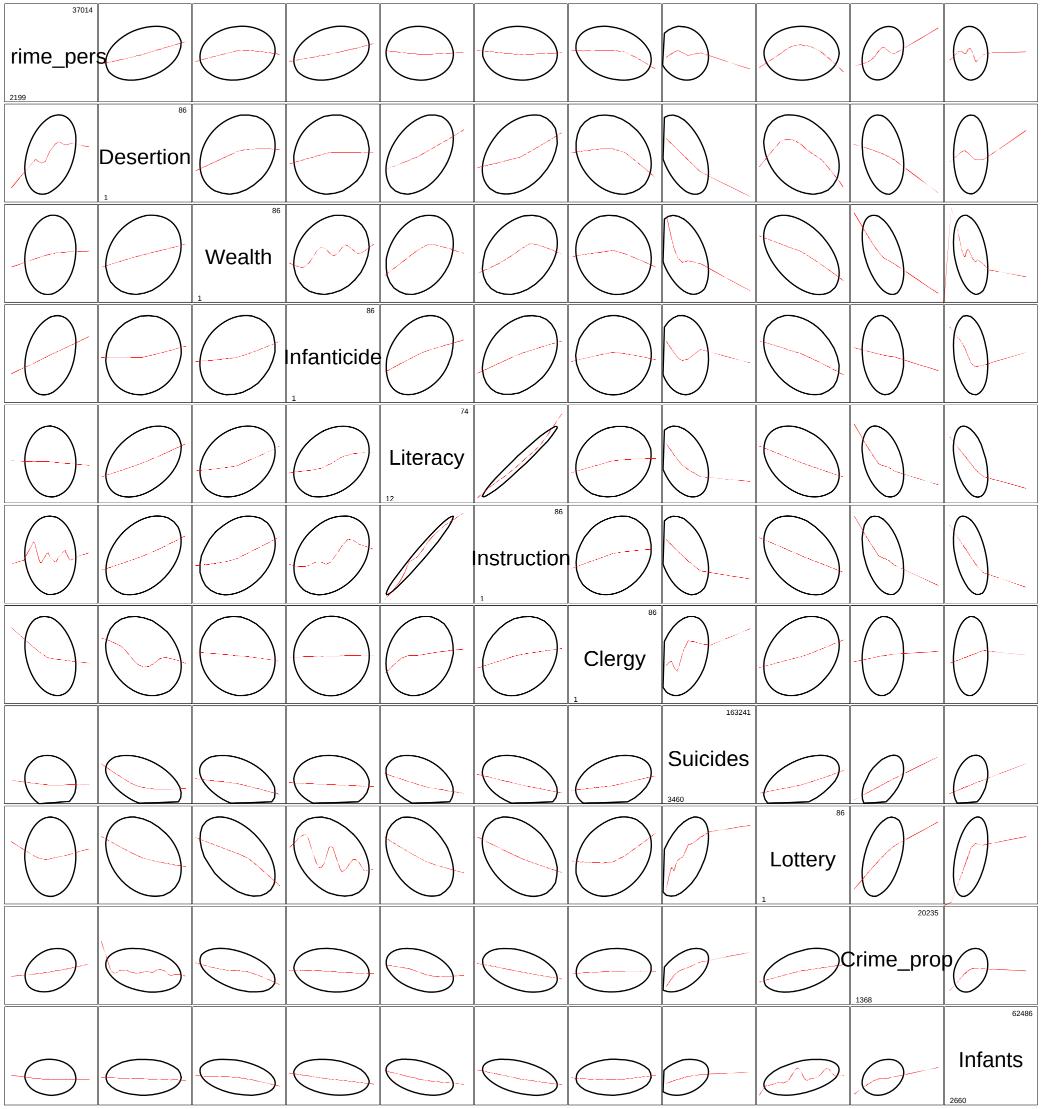


Guerry data— Variable ordering



Visual thinning: Minimal summaries for large data sets

Guerry data: schematic scatterplot matrix: 68% data ellipse + loess smooth



Multivariate analyses: Reduced rank displays

- Multivariate visualization techniques can show the statistical data in simple ways, using dimension reduction techniques.
  - Biplots - show variables and departments in space accounting for greatest variance
  - Canonical discriminant plots - show variables and departments in space accounting for greatest between-region variation
- Can try to show geographic location by color coding or other visual attributes.
  - Color code by region
  - Show data ellipse to summarize regions
- → **Data-centric displays**: The multivariate data is shown directly; geographic relations indirectly

Biplots

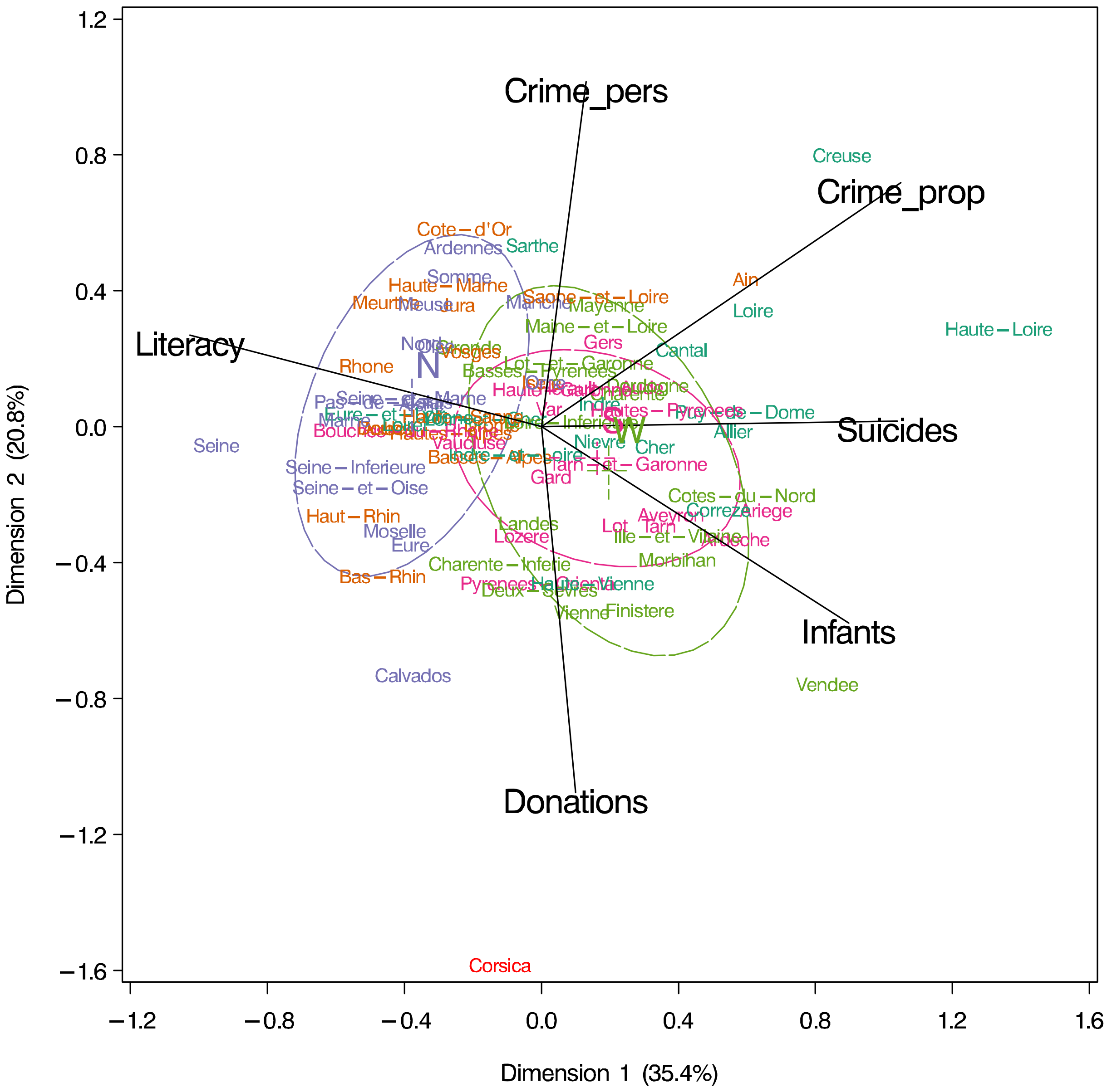
- Biplots represent both variables (attributes) and observations (departments) in the same plot— a low-rank (2D) approximation to a data matrix (Gabriel, 1971)

$$\boldsymbol{Y}^{\star} = \boldsymbol{Y} - Y_{..} \approx \boldsymbol{A}\boldsymbol{B}^{\mathsf{T}} = \sum_{k=1}^{d} \boldsymbol{a}_k \boldsymbol{b}_k^{\mathsf{T}}$$

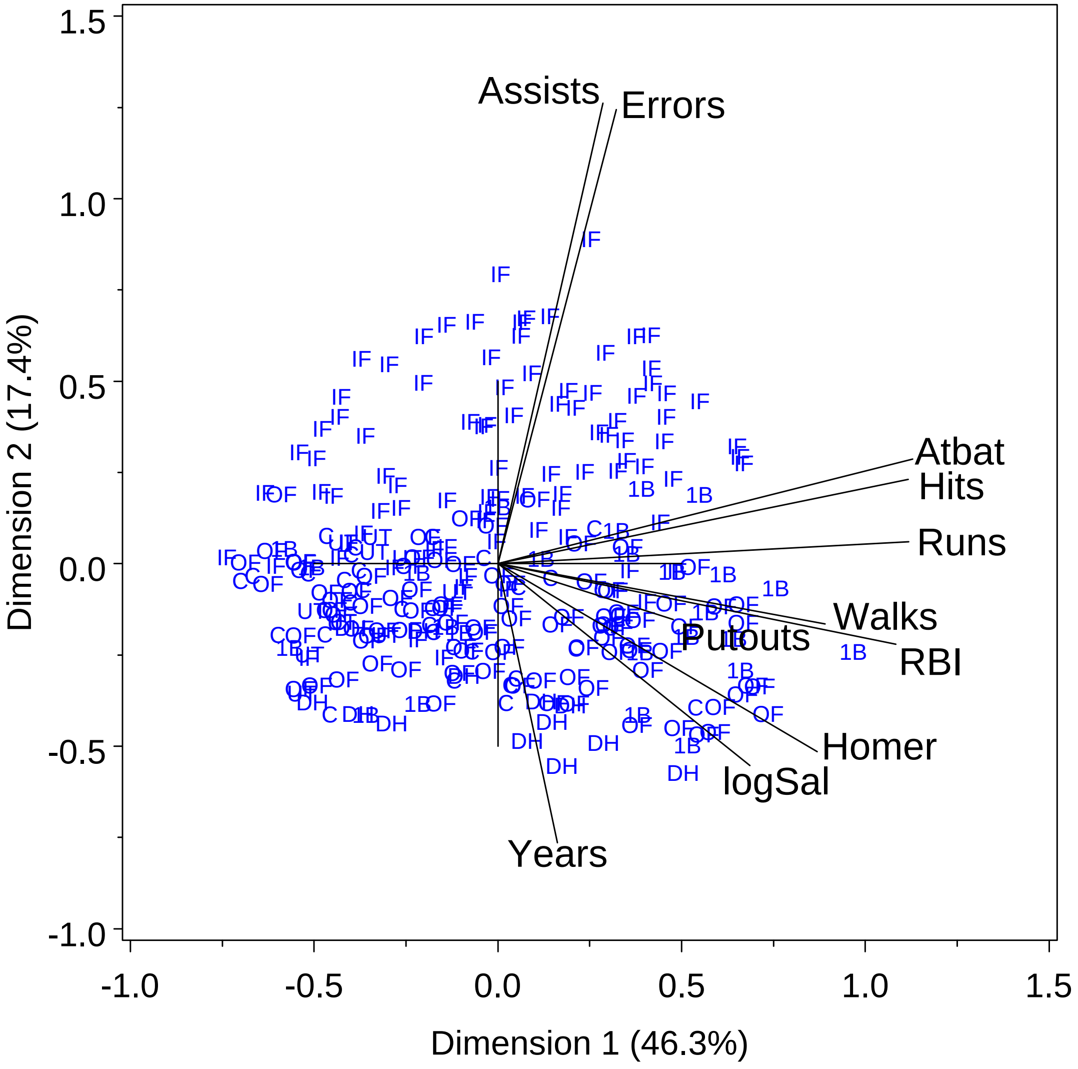
  - Variables are usually represented by vectors from origin (mean)
  - Observations are usually represented by points
  - Can show clusters of observations by data ellipses

- Properties:
  - Angles between vectors show correlations ($r \approx \cos(\theta)$)
  - Length of variable vectors $\sim$ % variance accounted for
  - $y_{ij} \approx \boldsymbol{a}_i^{\mathsf{T}} \boldsymbol{b}_j$: projection of observation on variable vector
  - Dimensions are uncorrelated overall (but not necessarily within group)

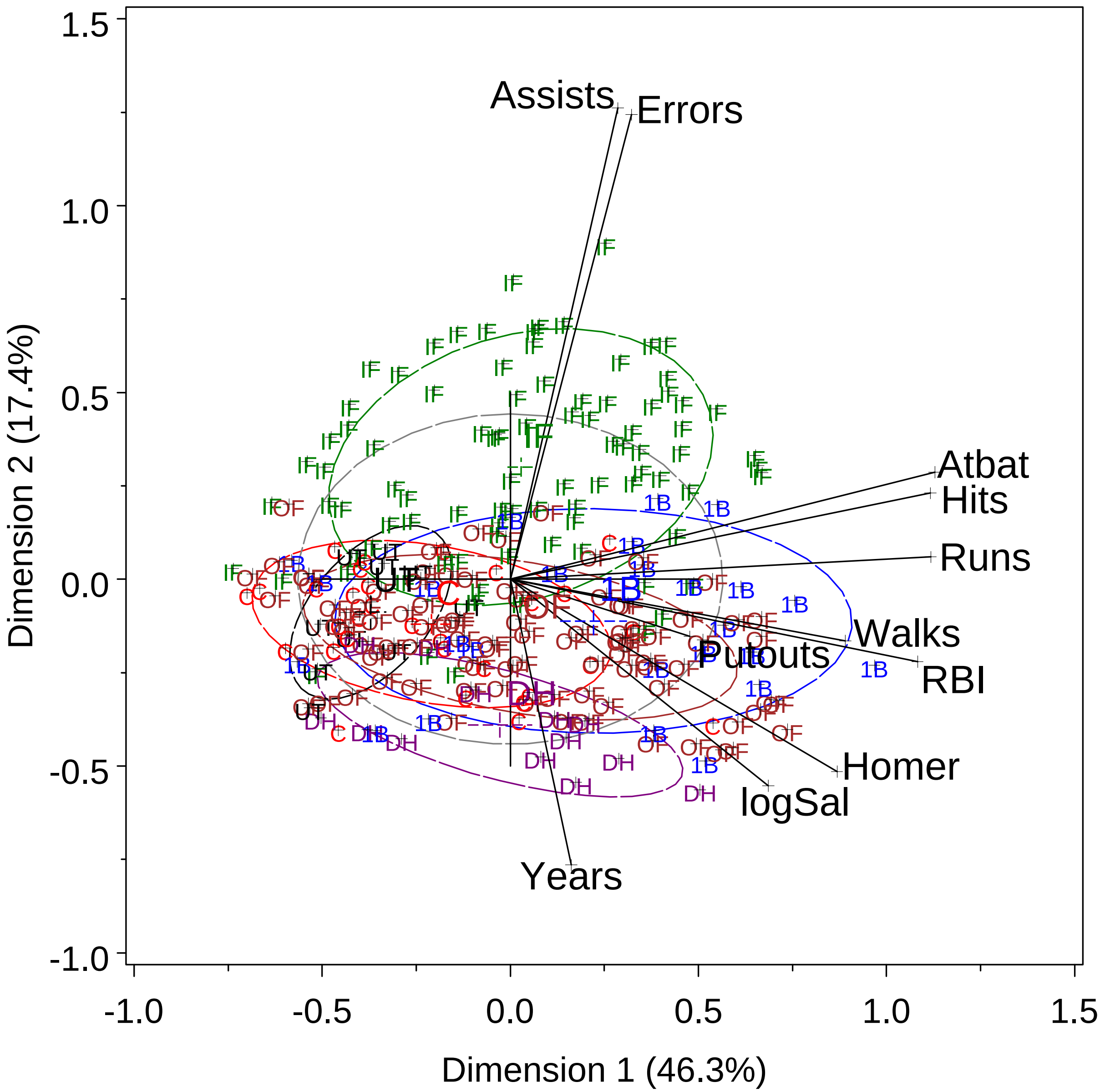
Biplots: Guerry data



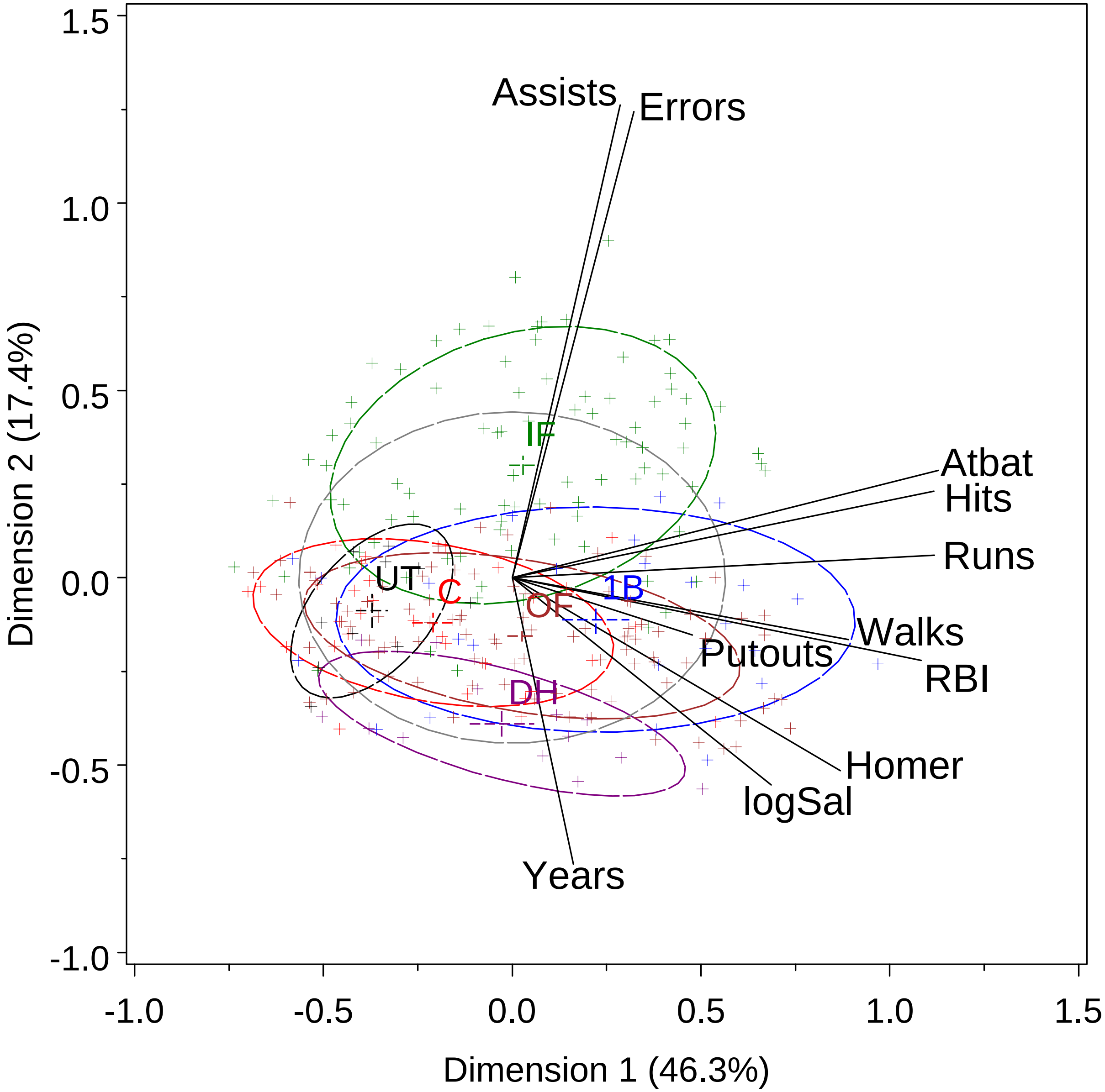
Biplots: Baseball data



Biplot of baseball data, players labeled by position



Biplot of baseball data, with data ellipses by position

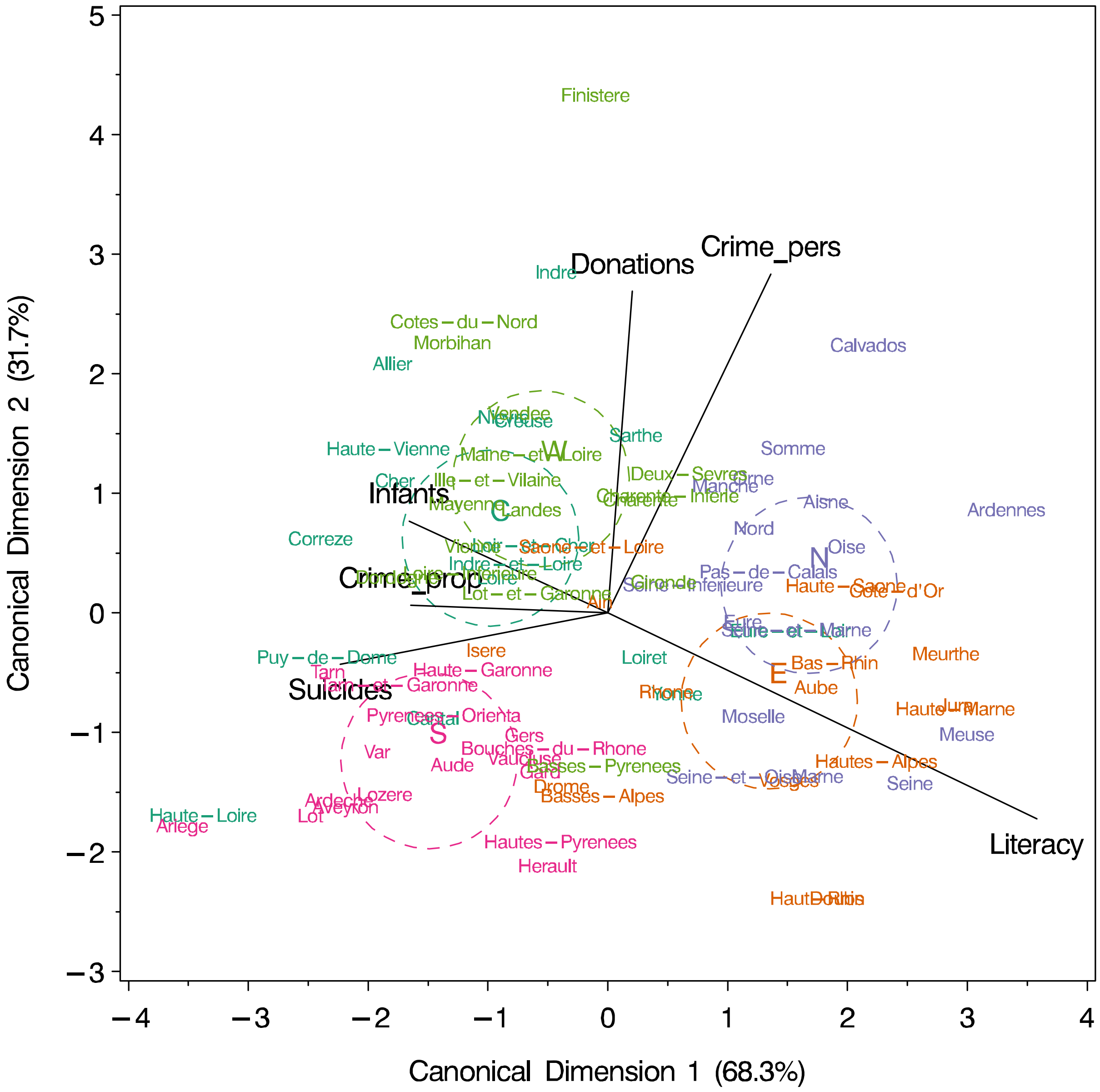


Biplot of baseball data, with data ellipses by position (thinned)

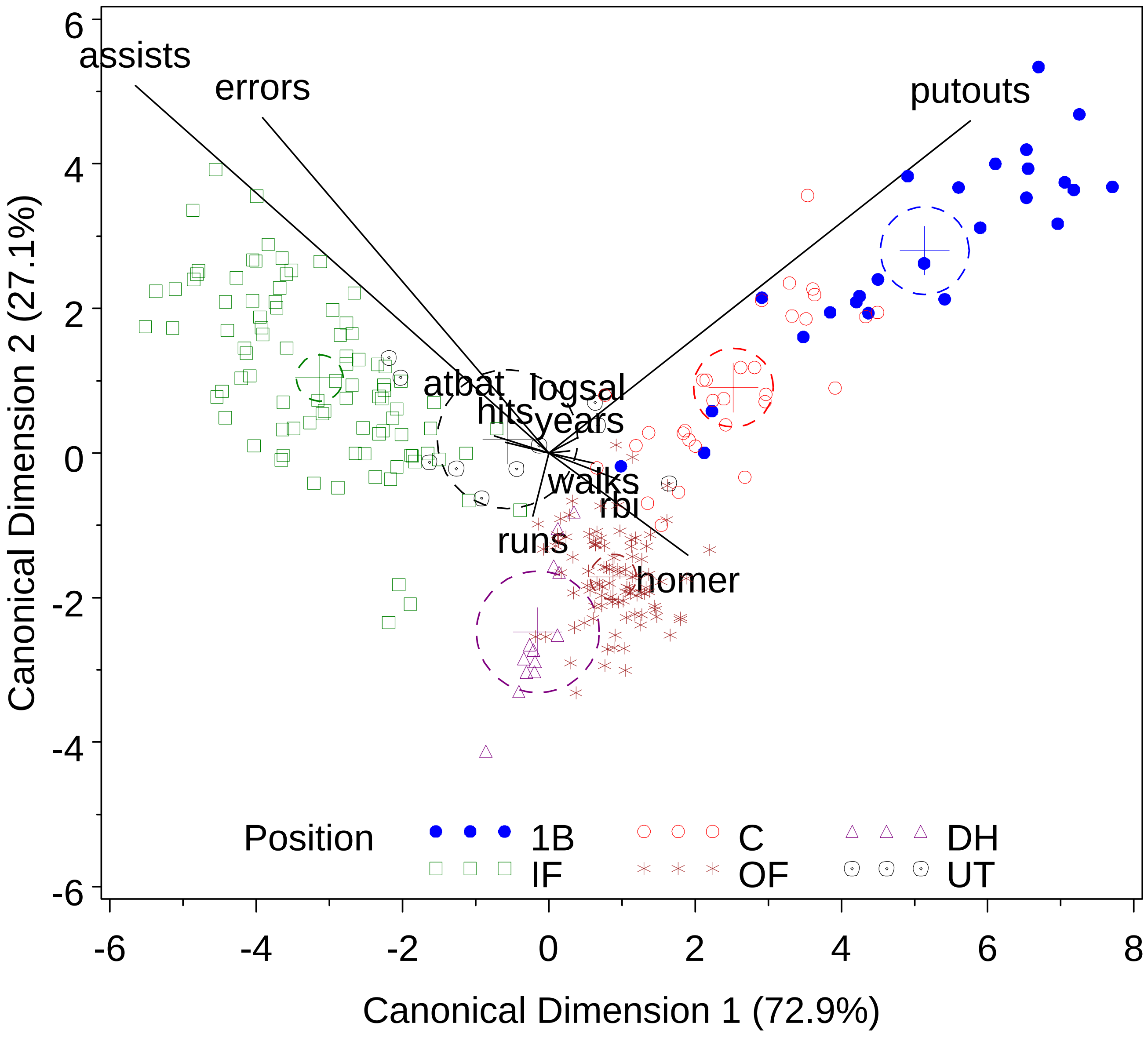
Canonical discriminant plots

- Project the variables into a low-rank (2D) space that maximally discrimates among regions (Friendly, 1991)
  - Visual summary of a MANOVA
  - Canonical dimensions are linear combinations of the variables with maximum univariate F-statistics.
  - Vectors from the origin (grand mean) for the observed variables show the correlations with the canonical dimensions
- Properties:
  - Canonical variates are uncorrelated
  - Circles of radius $\sqrt{\chi^2_2(1-\alpha)/n_i}$ give confidence regions for group means.
  - Variable vectors show how variables discriminate among groups
  - Lengths of variable vectors $\sim$ contribution to discrimination

Canonical discriminant plots: Guerry data, by Region



CDA plots: Baseball data, by player position



HE plots: Visualization for Multivariate Linear Models

- How are p responses, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p)$ related to q predictors, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q)$? (Friendly, 2004a)
  - MANOVA: $\boldsymbol{X} \sim$ discrete factors
  - MMRA: $\boldsymbol{X} \sim$ quantitative predictors
  - MANCOVA, response surface models,

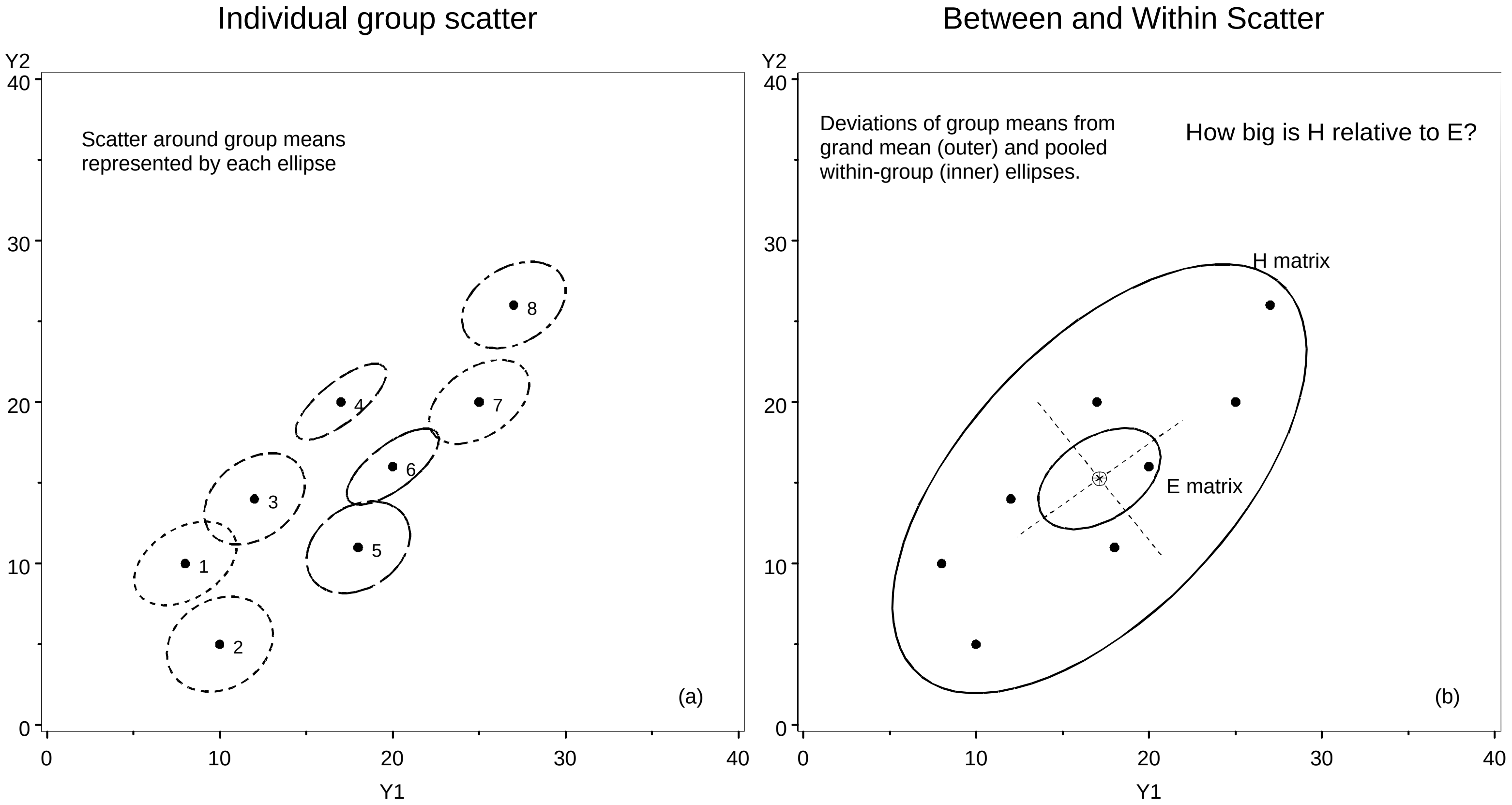
  } All the same MLM:

$$\underset{(n \times p)}{\boldsymbol{Y}} = \underset{(n \times q)}{\boldsymbol{X}} \underset{(q \times p)}{\boldsymbol{B}} + \underset{(n \times p)}{\boldsymbol{E}}$$

- Analogs of univariate tests:
  - Explained variation: $MS_H \longmapsto (p \times p)$ covariance matrix, $\boldsymbol{H}$
  - Residual variation: $MS_E \longmapsto (p \times p)$ covariance matrix, $\boldsymbol{E}$
  - Test statistics: $F \longmapsto |\boldsymbol{H} - \lambda \boldsymbol{E}| = 0 \mapsto \lambda_1, \lambda_2, \ldots \lambda_s$
- How big is $\boldsymbol{H}$ relative to $\boldsymbol{E}$?
  - Latent roots $\lambda_1, \lambda_2, \ldots \lambda_s$ measure the “size” of $\boldsymbol{H}$ relative to $\boldsymbol{E}$ in $s = \min(p, df_h)$ orthogonal directions.
  - Test statistics: Wilks’ Λ, Pillai trace, Hotelling-Lawley trace, Roy’s maximum root combine these into a single number

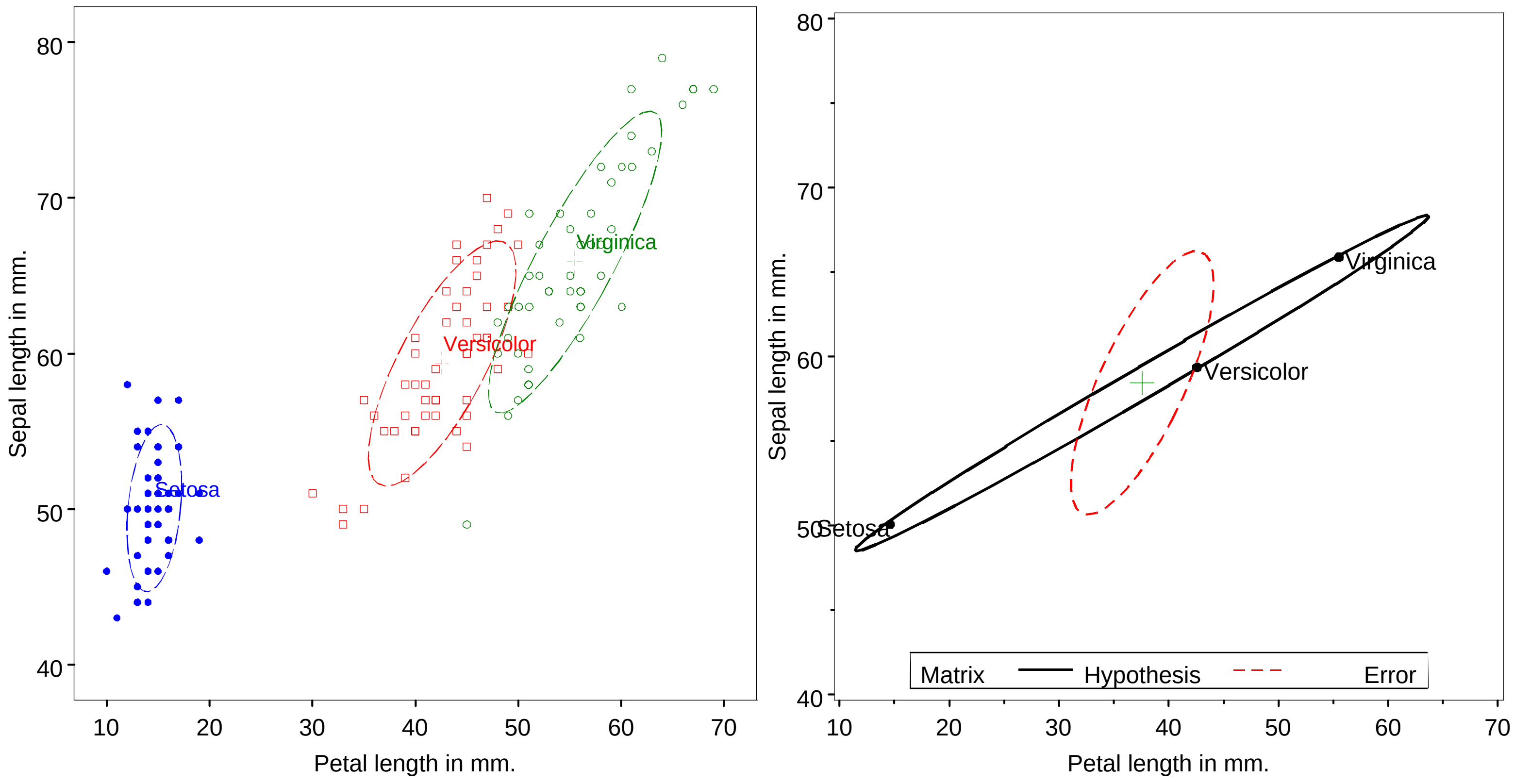
HE plots: Visualization for Multivariate Linear Models

- **HE plot**: for two response variables, $(\boldsymbol{y}_1, \boldsymbol{y}_2)$, plot a $\boldsymbol{H}$ ellipse and $\boldsymbol{E}$ ellipse
- **HE plot matrices**: For all p responses, plot an HE scatterplot matrix
- $\longrightarrow$ **Shows**: size, dimensionality, and effect-correlation of $\boldsymbol{H}$ relative to $\boldsymbol{E}$.



Essential ideas behind multivariate tests: (a) Data ellipses; (b) $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipses

Simple example: Iris data



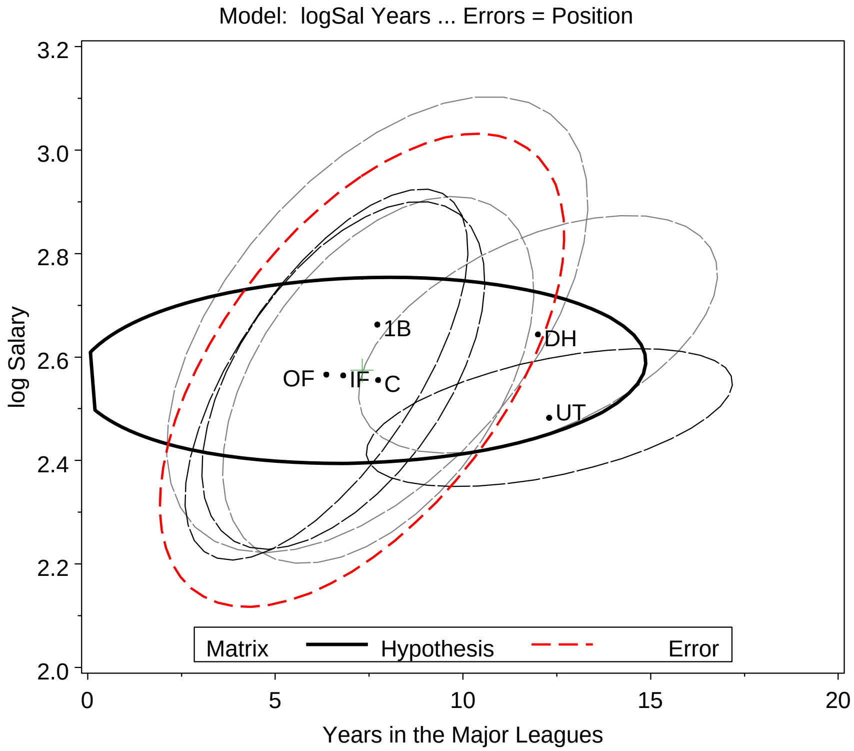
(a) Data ellipses and (b) $\boldsymbol{H}$ and $\boldsymbol{E}$ ellipses

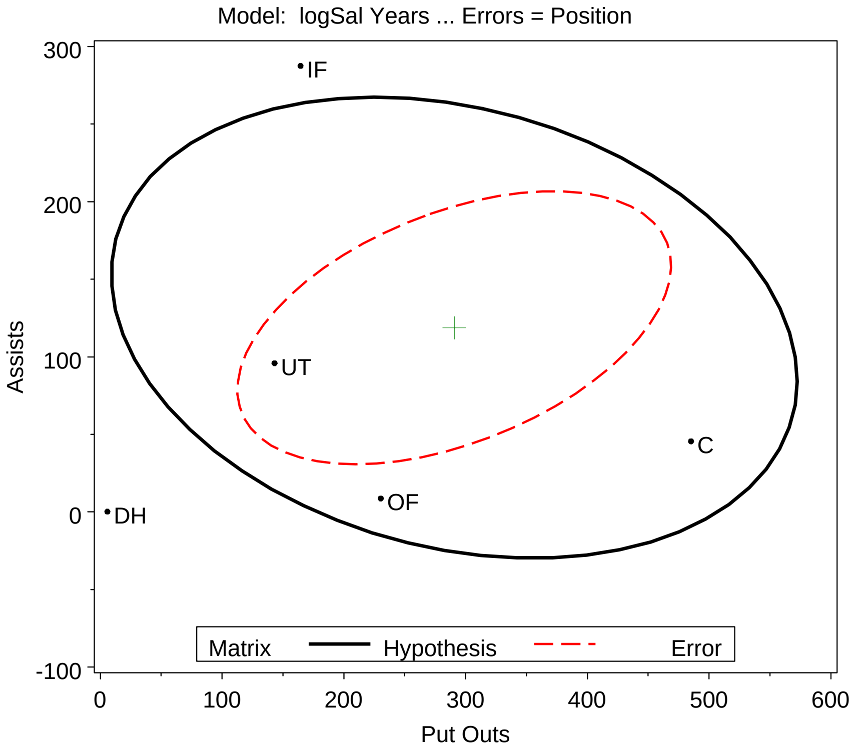
- $\boldsymbol{H}$ **ellipse**: Shows 2D covariation of *predicted* values (means)
- $\boldsymbol{E}$ **ellipse**: Shows 2D covariation of *residuals*
- **points**: show group means on both variables

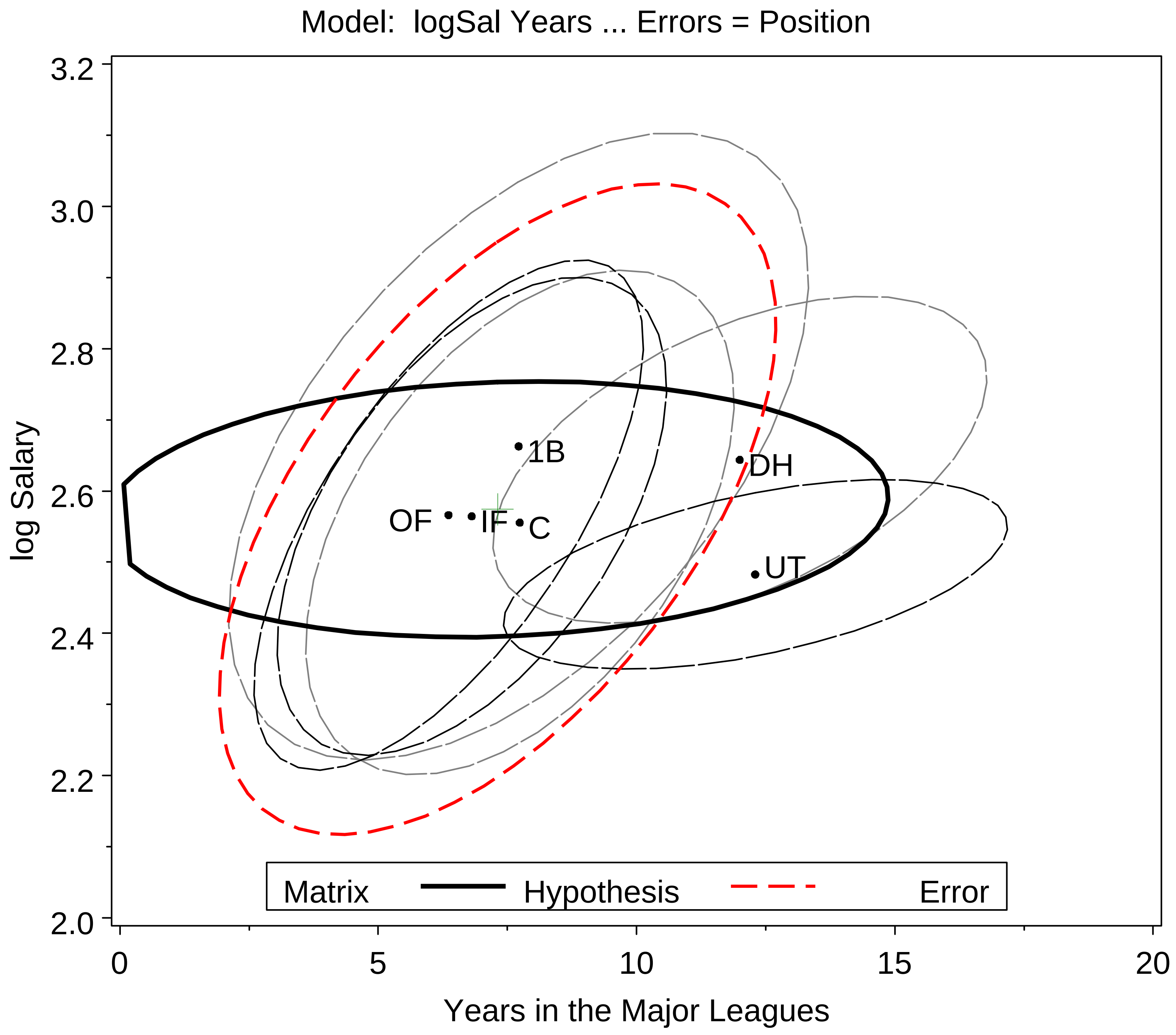
Baseball data: Variation by position

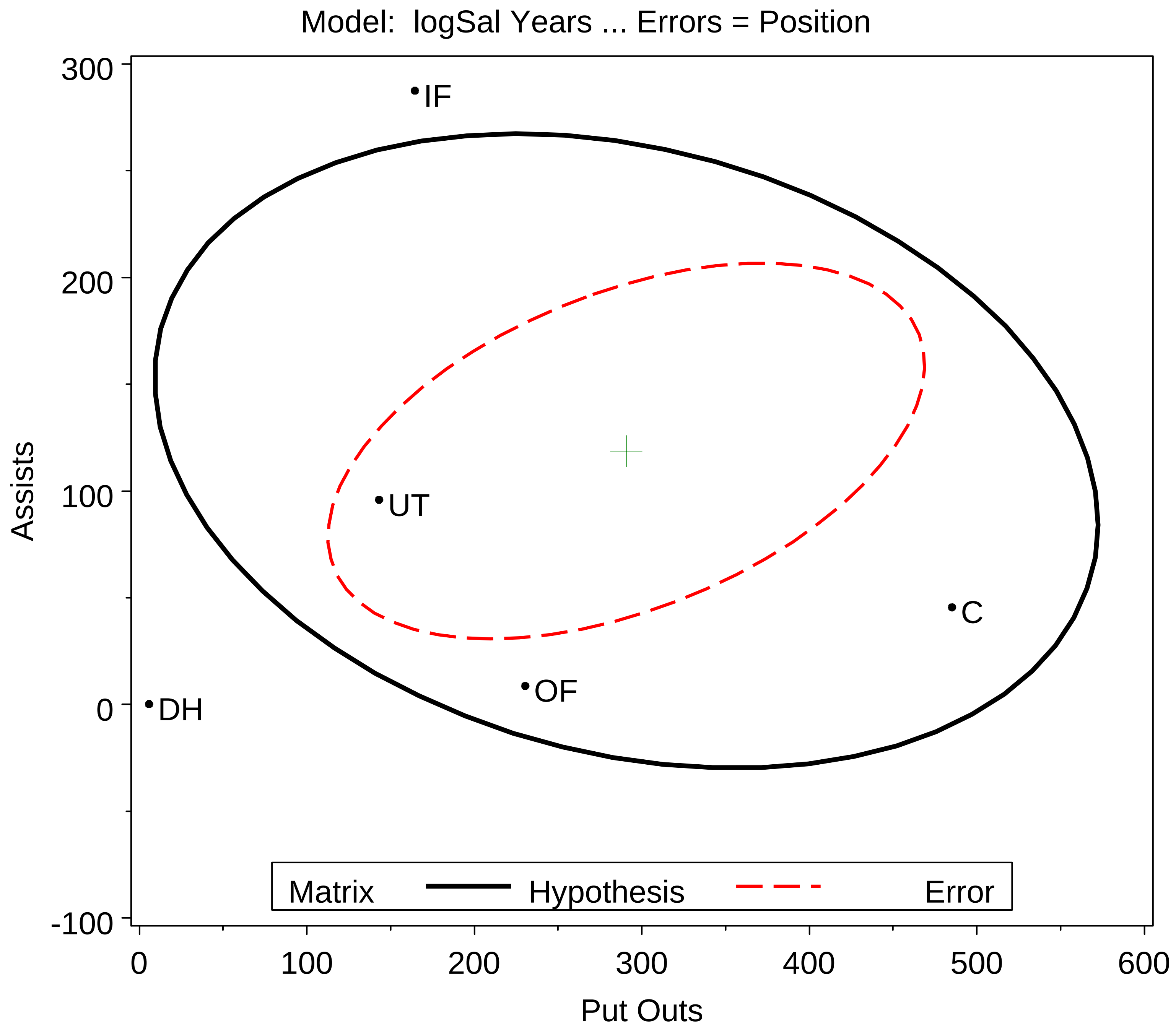
- How do relations among variables vary with player's position?
  - Fit MANOVA model,

    `(logSal Years Homer Runs Hits RBI Atbat Walks Putouts Assists Errors) = Position`
  - HE plots for selected pairs: (Years, logSal), (Putouts, Assists)






Model: logSal Years ... Errors = Position
log Salary
3.2
3.0
2.8
2.6
2.4
2.2
2.0
0
5
10
15
20
Years in the Major Leagues
1B
DH
OF
IF
C
UT
Matrix
Hypothesis
Error

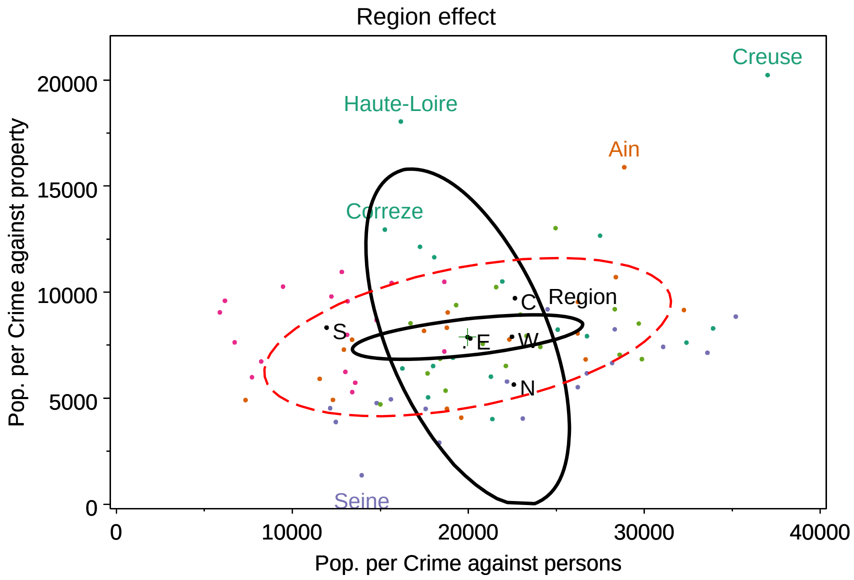

Model: logSal Years ... Errors = Position
Assists
Put Outs
300
200
100
0
-100
0
100
200
300
400
500
600
IF
UT
C
OF
DH
Matrix
Hypothesis
Error

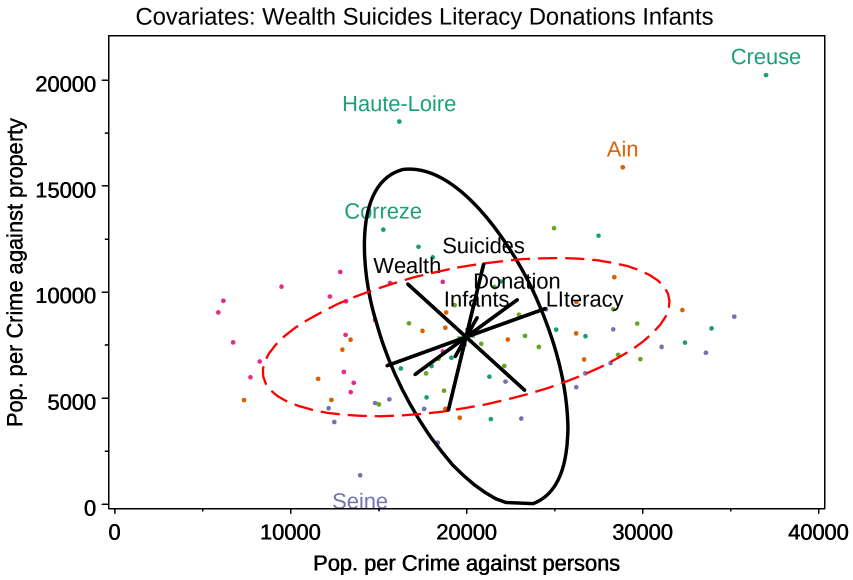
Guerry data: Predicting crime

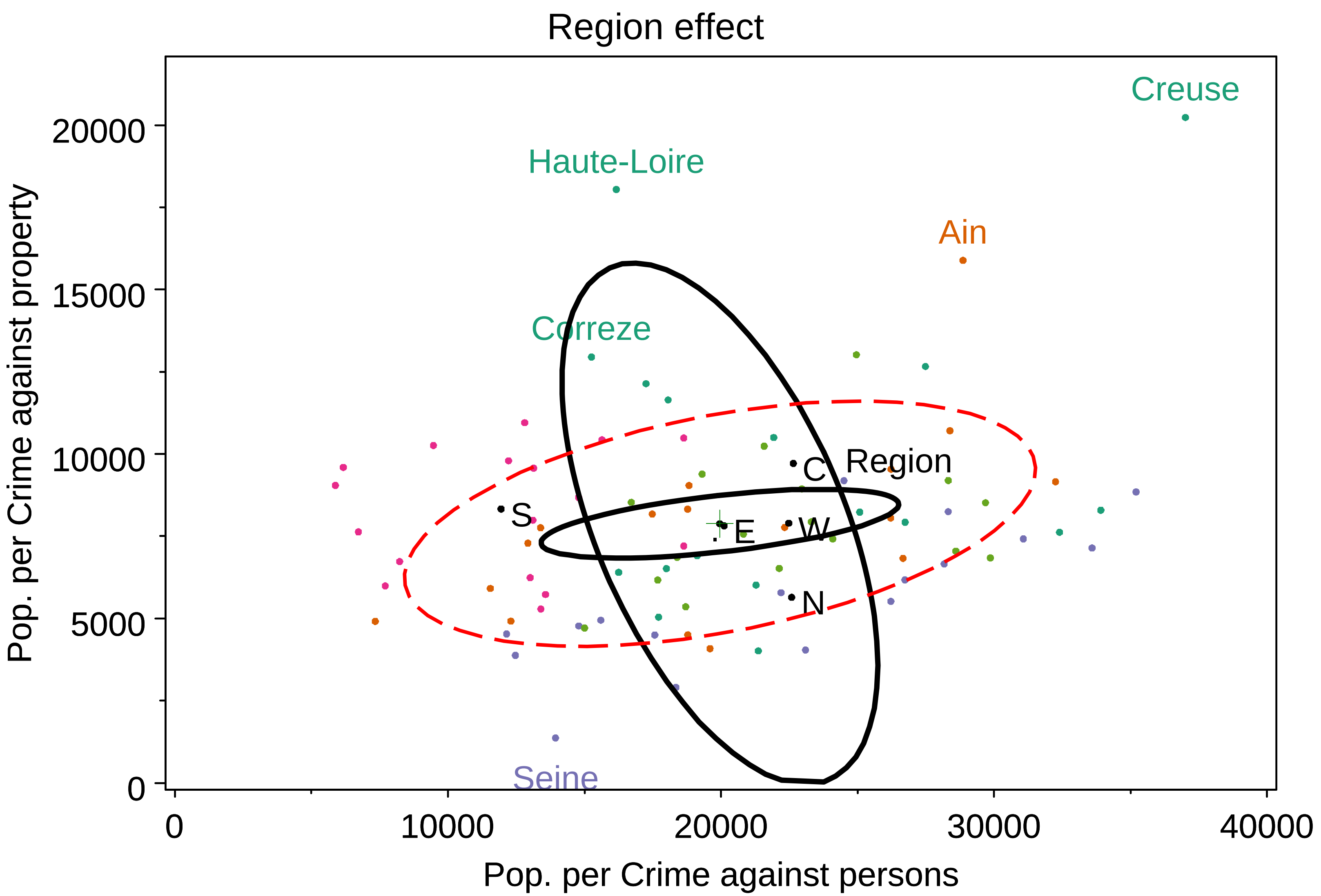
- How do rates of crime vary with other variables?
  - Fit MANCOVA model,

```
(Crime_pers Crime_prop) = Region + Wealth + Suicides
                        + Literacy + Donations + Infants
```

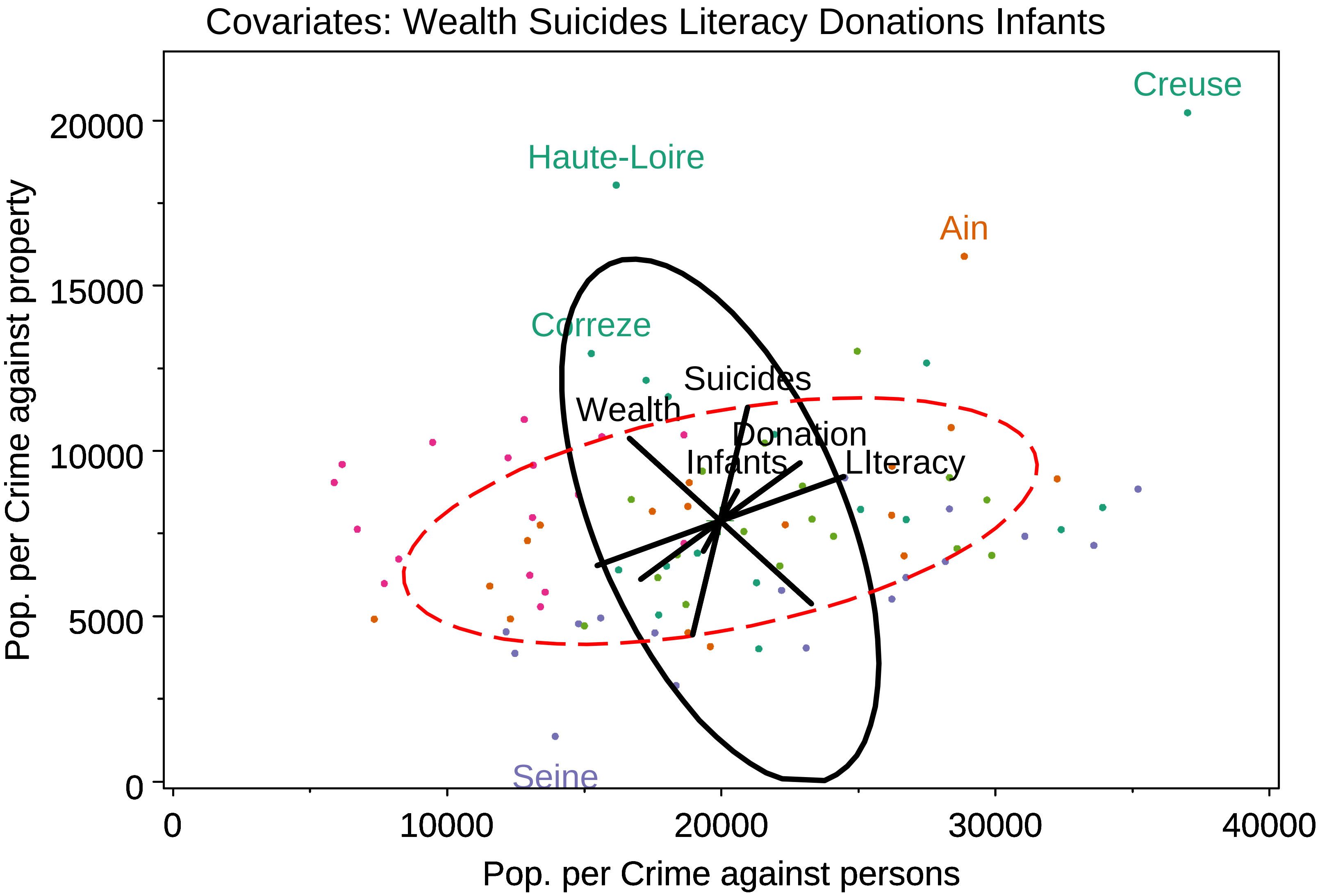
  - HE plots: Overall, plus for Region and covariate effects







- Overall: Predicted crimes against persons and property are negatively correlated
- Larger variation in crimes against property
- Region variation greater in crimes against persons



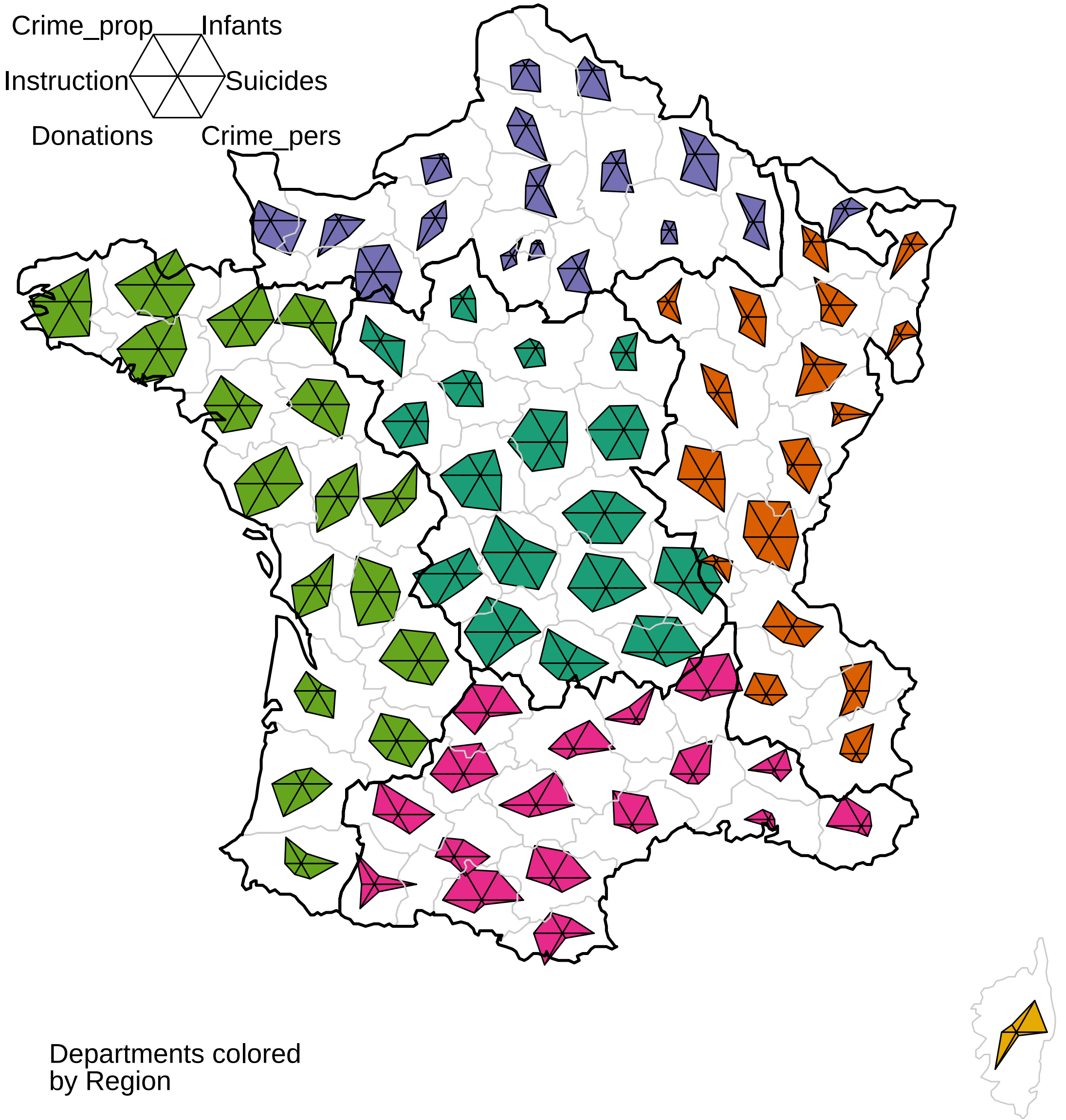
- Each quantitative variable (covariate) plots as a 1D ellipse (vector)
  - **Orientation**: relation of $\boldsymbol{x}_i$ to $\boldsymbol{y}_1, \boldsymbol{y}_2$
  - **Length**: strength of relation

Multivariate mapping: Map-centric displays

- How to generalize choropleth maps to many variables?
- **Star maps**: Show multivariate data on the map using star icons, variable $\sim$ length of ray
- **Reduced-rank RGB displays**: Factor analysis $\longrightarrow$ (F1, F2, F3) factor scores $\longmapsto$ (R, G, B) shading
- **PREFMAP (x, y) maps**: Fit data variables to (Long, Lat) map coordinates. Display variables as vectors in map coordinates.

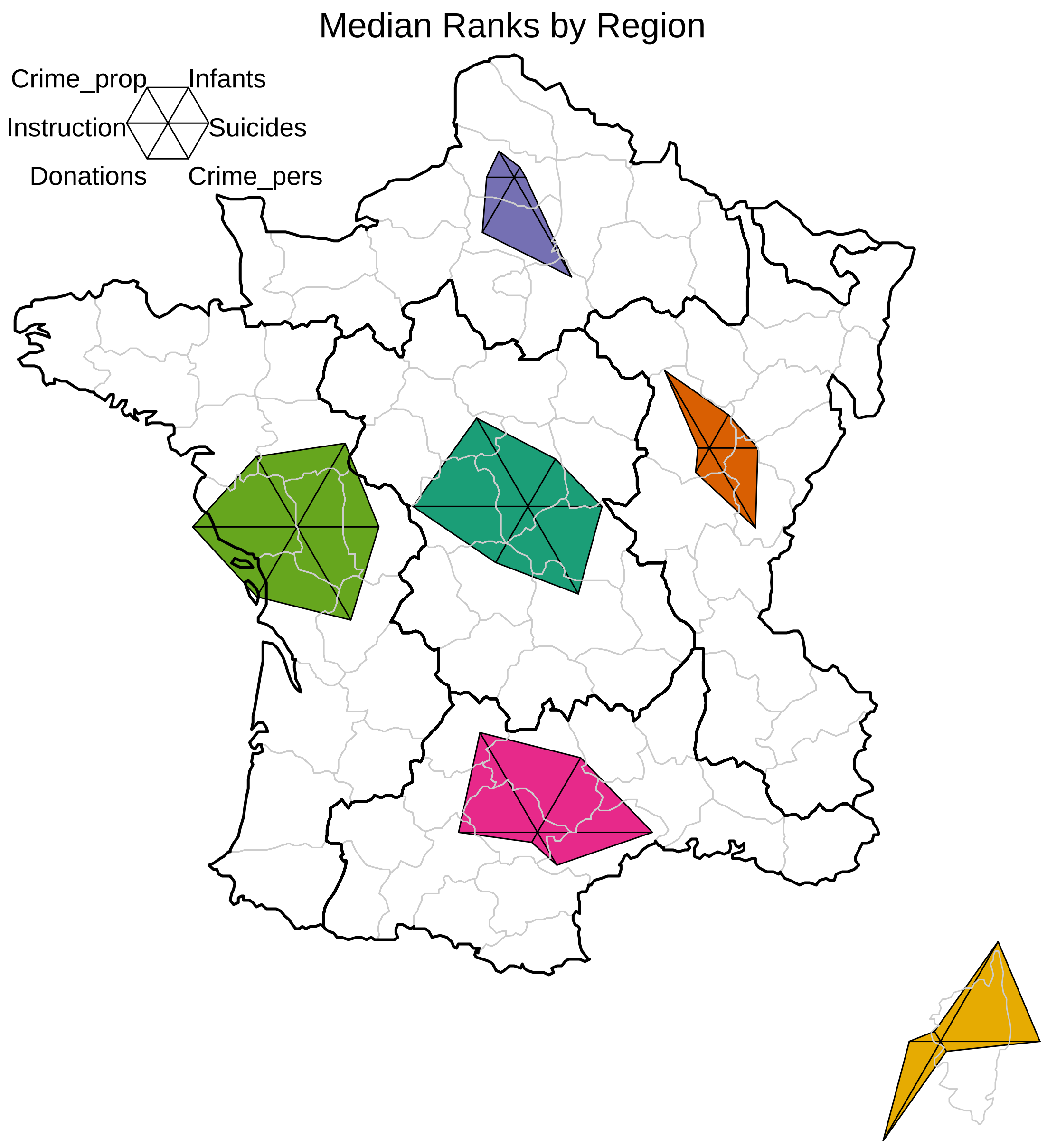
Star maps

Star map of Guerry Variables (Ranks)

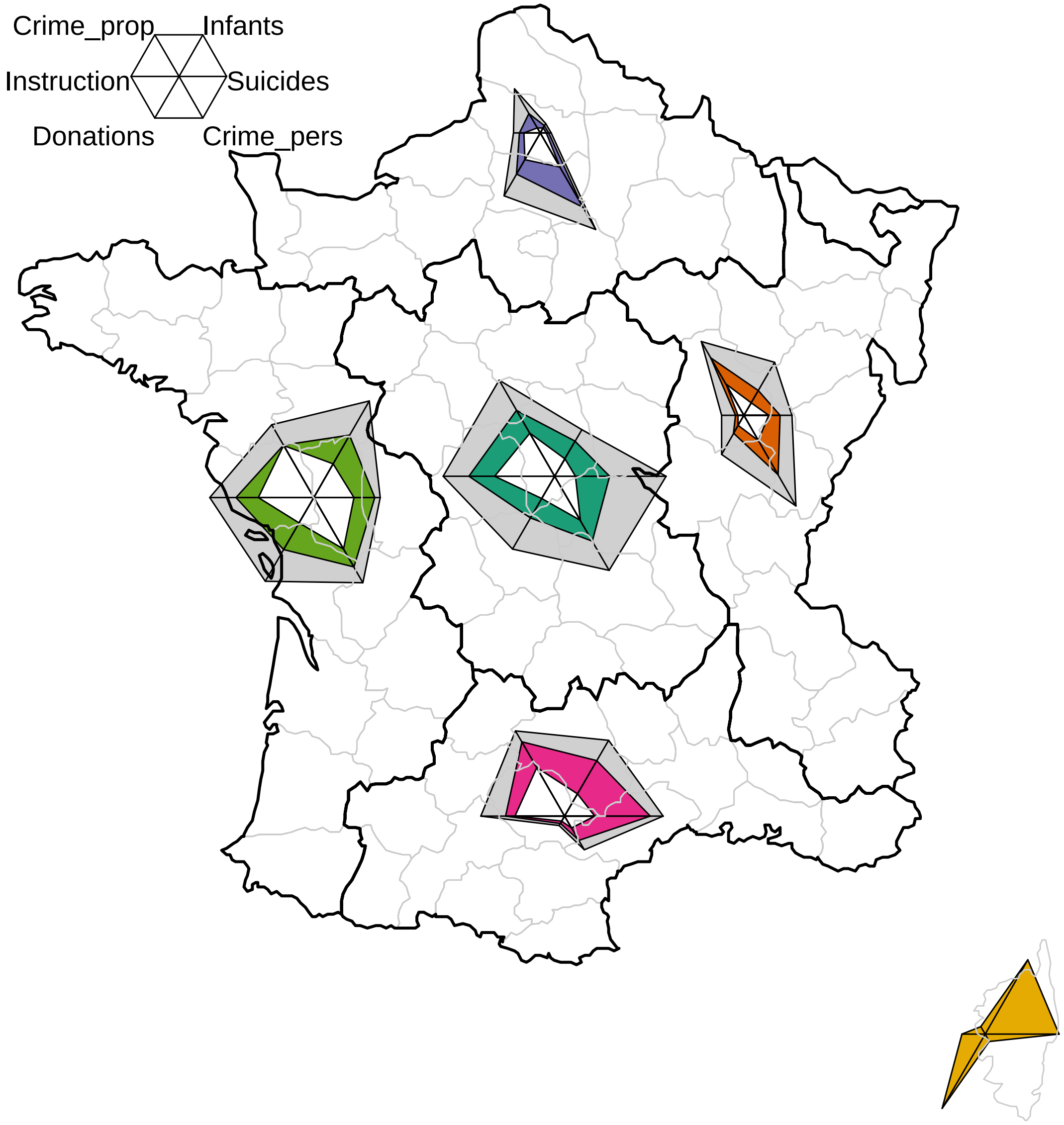


Departments colored
by Region

Star maps: Medians by region



Star maps: Multivariate boxplots by region



- stars for Q1, Median, Q3
- How to show unusual depts?

Reduced-rank color-coded displays

- Use dimension-reduction technique (PCA, Factor analysis, ...) to produce scores for observations (departments) on 3 dimensions (F_1, F_2, F_3)

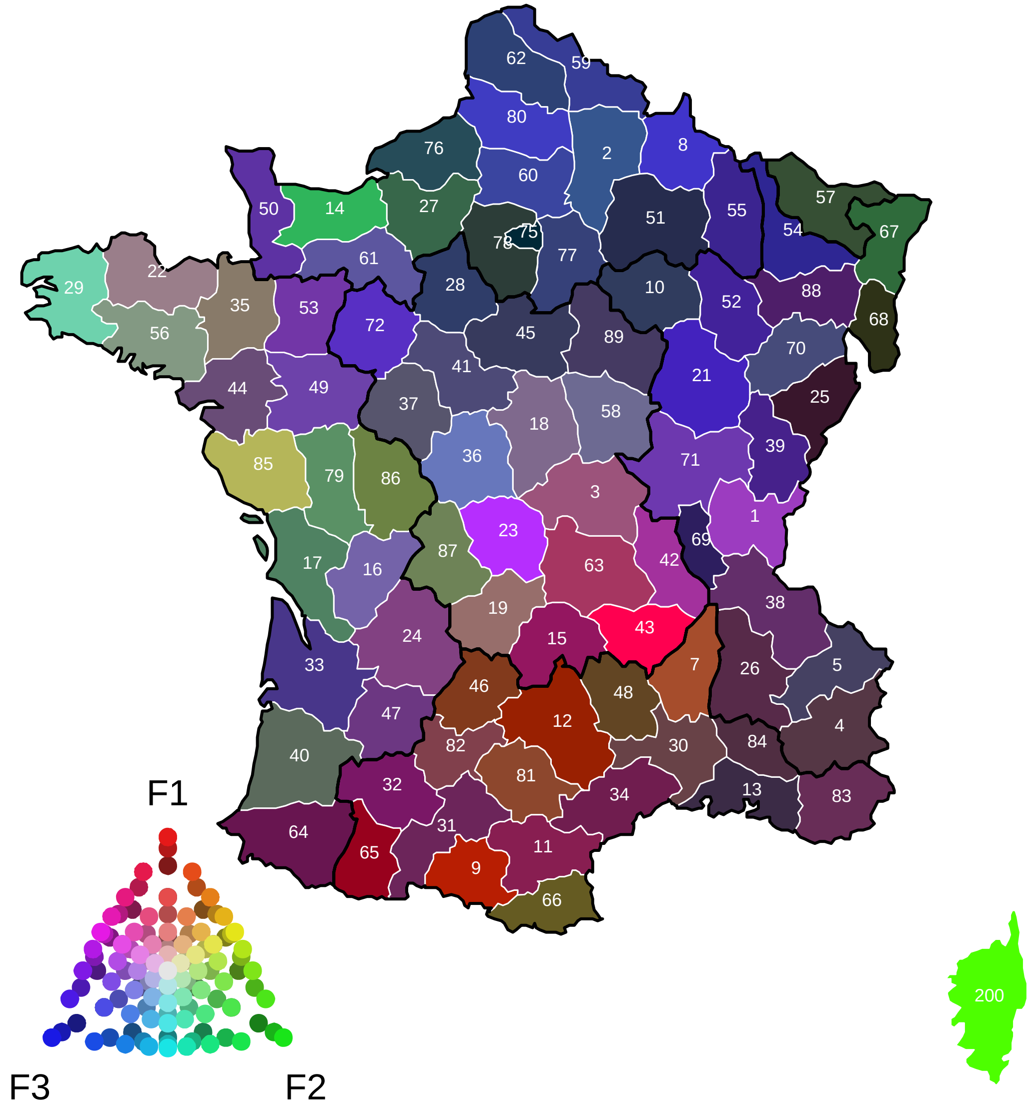
	Factor1	Factor2	Factor3
Variable	Civil society	Moral values	Crime
Pop per Crime against persons			0 97
Pop per Crime against property	0 75		0 39
Percent Read & Write	-0 72		
Pop per illegitimate birth	0 62	0 42	
Donations to the poor		0 89	
Pop per suicide	0 80		

- Scale $(F_1, F_2, F_3) \longrightarrow$ [0,1]
- Color mapping function, e.g., $\mathcal{C}(F_1, F_2, F_3) \mapsto$ rgb(F_i, F_j, F_k)

Reduced-rank color-coded displays

RGB 3-factor map: R=f1, G=f2, B=f3

Variables: Crime_pers Crime_prop Literacy Infants Donations Suicides



Summary and future directions

- **Guerry's challenge**: Understanding uncertainty in multivariate, spatial data
  - How visualize and understand relations among many variables?
  - How to relate these to geographic information?
- **Understanding multivariate variation**
  - Visual summaries (data ellipses, smoothings) can show statistical relations more clearly and effectively
  - Reduced-rank visualization methods can show simpler, approximate views, based on several criteria.
  - Multivariate statistical models need their own visualization methods, just beginning – HE plots as an example.
- **Understanding multivariate, spatial variation**
  - Multivariate visualizations applied to spatial data can be revealing, but still need work
  - Statistical methods for spatial data need to be extended to the multivariate setting.

References

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4), 316–324.

Friendly, M. (2004a). Graphical methods for the multivariate general linear model. *Journal of Computational and Graphical Statistics*. (submitted 2/17/04; resubmitted: 10/02/04).

Friendly, M. (2004b). Milestones in the history of data visualization: A case study in statistical historiography. In W. Gaul and C. Weihs, eds., *Studies in Classification, Data Analysis, and Knowledge Organization*. New York: Springer. (In press).

Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4), 509–539.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics*, 58(3), 453–467.

Guerry, A.-M. (1833). Essai sur la statistique morale de la France. Paris. English translation: Hugh P. Whitt and Victor W. Reinking, Lewiston, N.Y. : Edwin Mellen Press, c2002.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.