Supplementary material: Decision Heuristics for Comparison: How Good Are They?

Marcus Buckmann BUCKMANN@MPIB-BERLIN.MPG.DE
Özgür Şimşek OZGUR@MPIB-BERLIN.MPG.DE
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Editor: T.V. Guy, M. Kárný, D. Rios-Insua, D.H. Wolpert

Abstract

The first part of this document describes implementation details of the algorithms tested; the second part describes the 56 public data sets used in the empirical analysis.

1. Methods

Support vector machines. We used the support vector machine (SVM) implementation in the R package e1071 (Meyer et al., 2014). We tried both a linear and a Gaussian kernel. Both kernels have a cost parameter C. The Gaussian kernel in addition has a parameter γ that determines its spread. We selected both parameters using 10-fold cross-validated grid search. For the linear kernel, we evaluated the grid at $C = 2^p$, where p was set to odd integers ranging from -5 to 9. For the radial kernel, we evaluated the grid at $C = 2^p$, where p was set to odd integers ranging from -5 to 13, and at $\gamma = 2^p$, where p was set to even integers ranging from -16 to 8. We repeated our experiments with a nested grid search where we additionally evaluated a finer grid at 10 points ranging from 2^{p-1} to 2^{p+1}. The nested grid search did not improve the performance, so that the reported results are based on the simple grid search.

Random forest. We used the implementation of random forest in the R package randomForest (Liaw and Wiener, 2002). Typically, the only parameter tuned when using random forests is *mtry*, which specifies how many attributes should be randomly selected for consideration when splitting a branch (Hastie et al., 2009). We used 10-fold cross-validation to find the best value of *mtry* from the range $1, 2, ..., k$, where k is the number of available attributes. A second important parameter is *ntree*, which specifies the number of trees built. The default setting in the package (and used frequently in the literature) is 500. We repeated our experiments with *ntree* set to 500 and to 1,000; we did not see an observable difference. We report results with *ntree* set to 1,000. Random forests, in addition, have parameters that control how each individual tree is built. In the literature, these parameters are typically not tuned. We used the defaults in the package.

In his original paper on random forests, Breiman (2001) recommended setting *mtry* to $\lfloor\sqrt{k}\rfloor$, where k is the number of attributes. We tried this value as well. With few exceptions, the performance on individual learning curves either declined or remained the same. This

result is consistent with the earlier study by Fernández-Delgado et al. (2014). We report only the results obtained when setting *mtry* using cross-validation. For random forest regression we used the same implementation, set *ntree* to 1,000, and used cross-validation to find the best *mtry* parameter.

When learning naive Bayes, random forest, random forest regression and SVMs, we imposed the symmetry constraint by augmenting the training set with its mirror image (with respect to the direction of comparison). Each instance $\{\mathbf{\Delta x}^i_{AB}, u^i_{AB}\}$ is thus also framed as $\{\mathbf{\Delta x}^i_{BA}, u^i_{BA}\}$, where $u_{AB} = y^i_A - y^i_B$ for regressors and $u_{AB} = sgn(y^i_A - y^i_B)$ for classifiers.

Linear regression. We trained linear regression with elastic net regularization (Zou and Hastie, 2005), which contains both a ridge and a lasso penalty. The parameter α determines the relative proportion of ridge and lasso penalties, while the parameter λ controls the total penalty. We used the R package glmnet (Friedman et al., 2009) to train the model and set α and λ to the values that returned the minimum 10-fold cross-validation error in the training set. For α, we tested values from 0 to 1, in increments of 0.1, for λ we tested the values of the built-in search path. We imposed the symmetry constraint by setting the intercept to zero.

2. Data Sets

AFL Objects: 41 Australian Football League (AFL) games at the Melbourne Cricket Ground in 1993 and 1994. Criterion: attendance. Attributes: forecasted maximum temperature on the day of the game, total attendance at other AFL games in Melbourne and Geelong on the day of the game, total membership in the two clubs whose teams were playing, number of players in the top 50 who participated in the game, number of days since the earliest game of the season. Source: This data set was assembled by Rowan Todd and Mark McNaughton for a class project at the University of Queensland in a statistics course taught by Margaret Mackisack. The data sources were *The Football Bible '94* by Rex Hunt, *The Weekend Australian*, *Inside Football*, and *Football Record*. The data set is available from OzDASL data library (Smyth, 2011), where it is listed with the name *AFL Crowd Attendance at the MCG*.

Agriculture Objects: 29 groups of tropical subsistence cultivators. Criterion: agricultural intensity, defined as the proportion of time that each crop-fallow cycle is in the cropping phase. Attributes: population density, whether the group produces livestock, mean annual precipitation, length of dry season, soil type (normal, alluvial/hydromorphic), major staples of the group (root crops, cereal crops). Source: The data set is assembled by Turner et al. (1977) from earlier publications. It is electronically available from an online repository maintained by Winner, where it is listed with the name *population and other factors relating to agricultural intensity.*

Air Objects: 41 cities in the United States. Criterion: annual mean concentration of sulfur dioxide. Attributes: average annual temperature, number of manufacturing enterprises employing 20 or more workers, population, average annual wind speed, average annual rainfall, average number of days with rainfall per year. Source: The data were

gathered by Sokal and Rohlf (1981) from several publications of the United States government. The data set is reported in a book by Hand et al. (1994) with identifying number 26 and label *air pollution in US cities*.

Algae Objects: 340 samples from European rivers taken over a period of approximately one year. Criterion: density of algae type a. Attributes: concentrations of eight chemicals, season (fall, winter, spring, summer), river size (small, medium, large), fluid velocity (low, medium, high). Source: The data set is from the 1999 Computational Intelligence and Learning (COIL) competition. It is available from the UCI data repository (Bache and Lichman, 2013), where it is labeled *COIL 1999 competition data*.

Athlete Objects: 202 nationally-ranked athletes in Australia. Criterion: blood hemoglobin concentration. Attributes: body mass index, sum of skin folds, percent body fat, lean body mass, height, weight, sex, the sport the athlete competes in (basketball, field, gymnastics, netball, rowing, track 400m, swimming, sprint, tennis, water polo). Source: The data were collected by Telford and Cunningham (1991) at the Australian Institute of Sport. The data set is reported by Maindonald and Braun (2010) and is available from associated R package DAAG (Maindonald and Braun, 2013) with label *ais*.

Basketball Objects: 96 basketball players. Criterion: points scored per minute. Attributes: assists per minute, height, time played, age. Source: The data set is reported by Simonoff (1996) and is available from a website maintained by the author (Simonoff, 2015), where it is labeled *baskball.dat*.

Birthweight Objects: 189 newborns. Criterion: birth weight. Attributes: age of mother, weight of mother at last menstrual period, race (white, black, other), number of previous premature labors, number of physician visits during the first trimester, presence of uterine irritability, whether the mother smoked during pregnancy, whether the mother has a history of hypertension. Source: The data were collected at Baystate Medical Center in Springfield, Massachusetts in 1986 (Hosmer and Lemeshow, 2000). The data set is electronically available from R package MASS (Venables and Ripley, 2002; Ripley et al., 2013), where it is labeled *birthwt*.

Bodyfat Objects: 252 males. Criterion: percentage of body fat determined by underwater weighing. Attributes: age, weight, height, and various body circumference measurements: neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, wrist. Source: The data were collected by Penrose et al. (1985). The data set is available from StatLib (StatLib: Data, software and news from the statistics community) with label *bodyfat*.

Bone Objects: 42 male skeletons buried in coffins. Criterion: nitrogen content. Attributes: deposition time, depth of burial, age of the person, whether quicklime was added to the coffin at burial, whether skeleton was contaminated with oil, burial site (2 sites 130 km apart in northern England). Source: The data were collected by Jarvis (1997). The data set is electronically available from a data repository maintained by Winner, where it is listed with the name *nitrogen levels in skeletal bones of various ages and internment lengths*.

Car Objects: 93 passenger cars on sale in the United States in 1993. Criterion: sale price of the most basic version of the car. Attributes: city mileage, highway mileage, cylinders (3, 4, 5, 6, 8, rotary), engine size, maximum horsepower, engine revolutions per mile in highest gear, fuel tank capacity, passenger capacity, length, wheelbase, width, weight, rear seat room, luggage capacity, u-turn space, airbag (none, driver only, both driver and passenger), whether a manual transmission version is available, whether the manufacturer is from the United States, type of car (small, sporty, compact, midsize, large, van), drive-train type (rear, front, four-wheel drive). Source: The data set was assembled by Lock (1993) using information from *PACE New Car & Truck 1993 Buying Guide* and *Consumer Reports April 1993 Annual Auto Issue*. It is available from R package MASS (Venables and Ripley, 2002; Ripley et al., 2013) with label *Cars93*.

Cigarette Objects: 25 brands of cigarettes. Criterion: carbon monoxide emitted from the cigarette smoke. Attributes: weight, tar content, nicotine content. Source: The data were produced by the Federal Trade Commission. The data set is reported by Mendenhall and Sincich (1992). It is electronically available from the *Journal of Statistics Education* (McIntyre, 1994).

City Objects: 76 cities in Germany with more than 100,000 inhabitants. Criterion: population. Attributes: whether the city has a team in the major soccer league *Bundesliga*, whether the city is a state capital, whether the city was formerly in East Germany, whether the city is in the industrial belt, whether the abbreviation for the city on license plates is one-letter long, whether the city is on the intercity train line, whether the city hosted a trade fair in 2013, whether the city is the national capital, whether the city has a university. Source: This data set originally appeared in Gigerenzer and Goldstein (1996). For the current study, it was updated to reflect 2013 population data and attributes. Population data were obtained from the Federal Statistical Office (Das Statistische Bundesamt) based on the 2011 census and population density of revision 31.12.2012. Data on trade fairs were obtained from AUMA, the Association of the German Trade Fair Industry. The data set reflects only the trade fairs in AUMA category *international and national events*. The industrial belt is the region of Germany known as *Ruhrgebiet*. The intercity train line includes IC and ICE stops. Universities include *Universität*, *Institut für Technologie*, *Technische Universität*, and *Technische Hochschule*.

Contraception Objects: 210 localities in the world (most are United Nations members but includes areas like Hong Kong that are not independent countries). Criterion: percentage of unmarried women using a modern method of contraception. Attributes: annual population growth rate, per capita 2001 gross domestic product, percentage of females over the age of 15 who are economically active, population, expected number of live births per female in 2000, percentage of population that is urban in 2001. Source: The data set is reported by Weisberg (2005) who notes that the source of the data is the United Nations. It is electronically available from R package alr3 (Weisberg, 2011) where it is labeled *UN3*.

CPU Objects: 209 central processing units on the market in 1981–1984. Criterion: published performance on a benchmark mix relative to an IBM 370/158 Model 3. Attributes: cycle time, minimum main memory, maximum main memory, cache mem-

ory, minimum number of channels, maximum number of channels. Source: The data set was assembled by Ein-Dor and Feldmesser (1987) using information from *Computerworld* magazine. It is electronically available from R package MASS (Venables and Ripley, 2002; Ripley et al., 2013) with label *cpus*.

Crime Objects: 47 states of the United States. Criterion: crime rate in 1960. Attributes: percentage of males aged 14–24 in state population, indicator variable for a southern state, mean years of schooling of the population aged 25 years or older, per capita expenditure on police protection in 1960, per capita expenditure on police protection in 1959, labor force participation rate of civilian urban males in the age-group 14–24, number of males per 100 females, state population in 1960, percentage of nonwhites in the population, unemployment rate of urban males 14–24, unemployment rate of urban males 35–39, wealth (median value of transferable assets or family income), income inequality (percentage of families earning below half the median income), probability of imprisonment (ratio of number of commitments to number of offenses), average time served by offenders in state prisons before their first release. Source: The data set was assembled by Ehrlich (1973) from various publications of the United States government, including *Uniform Crime Reports* of the Federal Bureau of Investigation, United States Census, and *National Prison Statistics Bulletin*. Rounded data taken from Vandaele (1978) is electronically available from OzDASL (Smyth, 2011), where it is labeled *uscrime*.

Diamond Objects: 308 round diamond stones. Criterion: sale price. Attributes: weight in carats, color purity (D, E, F, G, H, I), clarity (internally flawless, very very slight inclusion 1, very very slight inclusion 2, very slight inclusion 1, very slight inclusion 2), certification (Gemmological Institute of America, International Gemmological Institute, Hoge Raad Voor Diamant). Source: The data set was assembled by Chu (2001) from advertisements in Singapore's *Business Times* edition of February 18, 2000. It is electronically available from R package Ecdat (Croissant, 2013).

Dropout Objects: 63 public high schools in Chicago. Criterion: dropout rate. Attributes: enrollment, attendance rate, parental involvement rate, percent limited-English students, percent low-income students, average class size, percent White students, percent Black students, percent Hispanic students, percent Asian students, percent minority teachers, average composite ACT score, IGAP scores: reading, math, science, social science, writing. Source: This prediction problem is from a study by Czerlinski et al. (1999). Their data sources are two articles in the February 1995 issue of *Chicago* magazine (Morton, 1995; Rodkin, 1995), where the authors note that their primary data source is Illinois State Board of Education's 1994 School Report Card.

Excavator Objects: 33 hydraulic excavators operating in the opencast mining industry in the United Kingdom. Criterion: annual maintenance cost. Attributes: weight, type of machine (front shovel, backacter), type of industry (opencast coal, opencast slate), company attitude to used oil analysis (regular use, not). Source: The data are from a study by Edwards et al. (2000). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *construction plant maintenance costs*.

Faculty Objects: 397 professors at a college in the United States. Criterion: academic year salary. Attributes: sex, rank (assistant professor, associate professor, full professor), year in service, number of years since phD was earned, department (theoretical, applied). Source: The data set is reported by Fox and Weisberg (2011b) and is electronically available from associated R package car (Fox and Weisberg, 2011a).

Fair Objects: 601 married individuals. Criterion: number of extramarital affairs in the past year. Attributes: sex, age, number of years in marriage, whether the individual has children, degree of religiosity (on a scale from 1 to 5), years of education, occupation (on a scale from 1 to 7, according to Hollingshead classification), marital happiness (on a scale from 1 to 5). Source: The data set was assembled by Fair (1978) using responses to a survey by *Psychology Today* in 1969. It is reported by Greene (2003) and is electronically available from R package Ecdat (Croissant, 2013).

Fuel Objects: 51 states and the District of Columbia of the United States. Criterion: per capita motor fuel consumption in 2001. Attributes: population, fuel tax rate, per capita income, miles of federal-aid primary highways, proportion of the population who are licensed drivers. Source: The data set is reported by Weisberg (2005) who notes that the source of the data is the Federal Highway Administration. The data set is available from R package alr3 (Weisberg, 2011) where it is labeled *Fuel2001*.

Galápagos Objects: 29 islands in the Galápagos archipelago. Criterion: number of plant species. Attributes: surface area, elevation, distance to the nearest island, surface area of the nearest island, distance from the center of the archipelago. Source: The data set was assembled by Johnson and Raven (1973). It is reported by Weisberg (2005) and is electronically available from associated R package alr3 (Weisberg, 2011) Notes: Elevations of six very small islands were not recorded in the original data set. These were taken from a version of the data set available from OzDASL (Smyth, 2011), labeled *Galápagos Island Species Data*, and where missing elevations were obtained from web searches and large-scale maps.

Highway Objects: 39 segments of highway in Minnesota. Criterion: accident rate. Attributes: segment length, average daily traffic count, truck volume as a percent of total volume, speed limit, number of lanes, lane width, shoulder width, number of signalized interchanges per mile, number of freeway-type interchanges per mile, number of access points per mile, highway type (federal interstate highway, principal arterial highway, major arterial, other). Source: The data set is reported by Weisberg (2005) who notes that the data were taken from an unpublished master's paper in civil engineering by Carl Hoffstedt. The data set is electronically available from R package alr3 (Weisberg, 2011).

Hitter Objects: 322 hitters in North American Major League Baseball. Criterion: annual salary at the beginning of the 1987 season. Attributes: 1986 performance: number of at bats, hits, home runs, runs scored, runs batted in, walks, putouts, assists, errors; career performance: number of at bats, hits, home runs, runs scores, runs batted in, walks; number of years in the major leagues; division at the end of the 1986 season (East, West); league at the end of the 1986 season (American, National); league at the beginning of the 1987 season (American, National). Source: The data set was prepared by the Statistical

Graphics Section of the American Statistical Association for the 1988 Annual Statistical Meetings and is available from StatLib (StatLib: Data, software and news from the statistics community). The version used in this work is from Fox (2008), includes corrections by Hoaglin and Velleman (1995), and is electronically available from a website maintained by Fox (2015).

Home Objects: 3281 homes sold in San Francisco. Criterion: sales price. Attributes: number of bedrooms, interior area of the property in squarefeet, lotsize of the property, year the property was built. Source: The data were reported in Adler (2010) and are available from associated R package nutshell (Adler, 2012) with label *sanfrancisco.home.sales*.

Homeless Objects: 50 cities in the United States. Criterion: rate of homelessness. Attributes: mean temperature, unemployment rate, percentage of inhabitants with incomes below the poverty line, vacancy rate, population, percentage of public housing, whether the city has rent control. Source: The data set was assembled by Tucker (1987) from Department of Housing and Urban Development's 1984 *Report to the Secretary on the Homeless and Emergency Shelters* and other sources.

Ice Objects: 30 four-week periods. Criterion: ice cream consumption per capita. Attributes: price of ice cream per pint, weekly family income, mean temperature. Source: The data set is reported by Hand et al. (1994) with identifying number 268 and name *ice cream consumption*. Their source is an article by Kadiyala (1970), who reports that the data are from a study by Hildreth and Lu (1960).

Infant Objects: 105 nations. Criterion: infant-mortality rate. Attributes: per-capita income, geographic location (Africa, Americas, Asia, Europe), whether the country exports oil. Source: Rates of infant mortality were obtained by Leinhardt and Wasserman (1979) from the editorial section of the *New York Times* (Crittenden, September 28 1975). The data set is reported by Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Jet Objects: 22 jet fighter aircraft of the United States Navy and Air Force. Criterion: First flight date, in months after January 1940. Attributes: Specific power, flight range factor, payload as a fraction of gross weight, sustained load factor, whether the aircraft can land on a carrier. Source: The data set was assembled by Stanley and Miller (1979). It is reported in a book by Hand et al. (1994) with identifying number 110.

Lake Objects: 69 world lakes. Criterion: number of known crustacean zooplankton species present. Attributes: surface area, maximum depth, mean depth, specific conductance, elevation, latitude, longitude, distance to nearest lake, number of lakes within 20 km, rate of photosynthesis. Source: The data set is reported by Weisberg (2005) who notes that the data were provided by Dodson and discussed in part in Dodson (1992). The data set is electronically available from R package alr3 (Weisberg, 2011).

Land Objects: 67 counties in Minnesota. Criterion: rent per acre paid in 1977 for agricultural land planted in alfalfa. Attributes: average rent for all tillable land, density of dairy cows, proportion of pasture land, whether liming is required to grow alfalfa. Source:

The data set is reported by Weisberg (2005) who notes that the data were collected by Douglas Tiffany. The data set is electronically available from R package alr3 (Weisberg, 2011) where it is labeled *landrent*.

Lung Objects: 654 children. Criterion: forced expiratory volume in liters. Attributes: age in years, height in inches, gender, exposure to smoking. Source: The data were collected by Tager et al. (1979). The data set is reported in Ekstrom and Sørensen (2010) and is electronically available from associated R package isdals (Ekstrom and Sorensen, 2014) where it is labeled *fev*.

Mammal Objects: 62 mammal species. Criterion: average daily sleep. Attributes: body weight, brain weight, maximum life span, gestation time, predation index, sleep exposure index, overall danger index. Source: The data are from a study by Allison and Cicchetti (1976). The data set is available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *sleep*.

Manpower Objects: 17 naval hospitals of the United States around the world. Criterion: monthly man-hours. Attributes: average daily patient load, monthly X-ray exposures, monthly occupied bed days, eligible population in the area, average length of stay by a patient. Source: The data were obtained by Myers (1990) from a publication of the United States Navy (1979). The data set is electronically available from R package genridge (Friendly, 2012). A similar data set is reported in a book by Hand et al. (1994) with identifying number 269 and label *hospital data*.

Mileage Objects: 398 cars built in 1970–1982. Criterion: mileage. Attributes: number of cylinders, engine displacement, horsepower, vehicle weight, time to accelerate from 0 to 60 mph, model year, origin (American, European, Japanese). Source: The data set was prepared by the Committee on Statistical Graphics of the American Statistical Association for its Second Exposition of Statistical Graphics Technology, held in conjunction with the Annual Meetings in Toronto, August 15–18, 1983. It is electronically available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *cars*. The version used in the current work is from the UCI Machine Learning Repository (Bache and Lichman, 2013), named *Auto+MPG*, in which 8 of the original cars were removed because their mileage values were missing.

Mine Objects: 44 coal mines in the Appalachian region of western Virginia. Criterion: number of fractures in upper seams of coal mines. Attributes: inner burden thickness, percent extraction of the lower previously mined seam, lower seam height, duration of operation. Source: The data set is reported by Montgomery et al. (2001) and is electronically available from associated R package mpg (Braun, 2012) where it is labeled *p13.7*.

Monet Objects: 430 sales of paintings by Monet. Criterion: sale price. Attributes: height of the painting, width of the painting, whether the painting is signed, auction house where sale took place. Source: The data set is reported by Greene (2003). It is electronically available from a website maintained by the author (Greene), where it is labeled *data on sales of Monet paintings*.

Mortality Objects: 60 metropolitan areas in the United States. Criterion: mortality rate. Attributes: average annual precipitation, average January temperature, average July temperature, percent population aged 65 or older, average household size, median school years completed by those over 22, percent housing units that are sound and with all facilities, humidity, population density in urbanized areas, percent nonwhite population in urbanized areas, percent employed in white collar occupations, percentage of families with income less than $3000, relative hydrocarbon pollution potential, relative nitric oxides pollution potential, relative sulfur dioxide pollution potential, annual average relative humidity. Source: The data set was assembled by McDonald and Schwing (1973). It is electronically available from StatLib (StatLib: Data, software and news from the statistics community), where it is labeled *pollution*.

Movie Objects: 62 movies. Criterion: first-run box office in the United States. Attributes: production budget, index of star poser, whether the movie is a sequel, indicator for an action film, indicator for comedy, indicator for animation, indicator for horror, MPAA rating(G, PG, PG13, R), trailer views at traileraddict.com, number of message board comments at comingsoon.net, attention at fandango.com, percentage of Fandango votes for "can't wait to see". Source: The data set is reported by Greene (2003) and is electronically available from a website maintained by the author (Greene), where it is labeled *movie buzz data*.

Mussel Objects: 44 rivers in eastern United States. Criterion: number of freshwater mussel species. Attributes: area of drainage basins, amount of dissolved solids, nitrate concentration, hydronium concentration, number of intervening rivers to four major species-source river systems: Alabama-Coosa, Apalachicola, Savannah, and St. Lawrence. Source: The data are from an article by Sepkoski and Rex (1974). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *freshwater mussel species in US Rivers*.

Obesity Objects: 136 children. Criterion: somatotype (a scale of body type, ranging from 1, very thin, to 7, obese). Attributes: sex, body measurements at ages 2, 9, and 18: height, weight, leg circumference, strength. Source: The data were collected by Tuddenham and Snyder (1954) on children born in Berkeley, California, between January 1928 and June 1929. The data set is reported by Weisberg (2005) and is electronically available from associated R package alr3 (Weisberg, 2011) where it is labeled *BGSall*.

Occupation Objects: 36 occupations. Criterion: prestige rating of the National Opinion Research Center (NORC). Attributes: suicide rate among males aged 20–64, median income, median number of school years completed. Source: The data set was assembled by Labovitz (1970) using data from the U.S. Census of 1950 and prestige rankings obtained by NORC in its 1947 survey. It is reported in a book by Hand et al. (1994) with identifying number 490 and label *prestige, income, education, and suicide rates for 36 occupations.* Notes: For some occupations, median number of school years completed is reported as 16+. These values were treated as 16 in the analysis.

Oxidant Objects: 30 summer days in Los Angeles, California. Criterion: maximum level of an oxidant. Attributes: morning averages of four meteorological variables: wind

speed, temperature, humidity, insolation. Source: The data set is reported by Rice (1995) (pp. 567–570), who notes that the data were collected by the Los Angeles Pollution Control District.

Pinot Objects: 38 samples of Pinot Noir wine. Criterion: quality. Attributes: clarity, aroma, body, flavor, oakiness, region. Source: The data set is reported by Montgomery et al. (2001) and is electronically available from associated R package MPV (Braun, 2012), where it is labeled *table.b11*.

Pitcher Objects: 206 pitchers in North American Major League Baseball. Criterion: annual salary at the beginning of the 1987 season. Attributes: 1986 performance: wins, losses, earned run average, game appearances, innings pitched, games saved; career performance: wins, losses, earned run average, game appearances, innings pitched, games saved; years in major leagues; league at the end of 1986 (American, National); league at the beginning of the 1987 season (American, National). Source: The data set was prepared by the Statistical Graphics Section of the American Statistical Association for the 1988 Annual Statistical Meetings and is available from StatLib (StatLib: Data, software and news from the statistics community). The version used in this work is from Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Prefecture Objects: 45 prefectures in Japan. Criterion: number of emigrants to Pacific Northwest in 1911–1912 from the prefecture (per million of the prefecture's population). Attributes: percentage of land cultivated by tenant farmlands, change in ratio of tenant farmlands between 1883 and 1907, average area of arable land per farm, number of government contracted laborers sent to Hawaii, whether any of the 18 pioneer Japanese immigrants to the Pacific Northwest were from the prefecture. Source: The data are from an article by Murayama (1991). The data set is electronically available from an online repository maintained by Winner, where the data set is described as *Japanese emigration to Pacific Northwest 1880–1915*.

Reactor Objects: 32 light water reactors constructed in the United States in the late 1960s and early 1970s. Criterion: construction cost. Attributes: date on which the construction permit was issued (measured in years since January 1, 1900), time between application for and issue of the construction permit, time between issue of operating license and construction permit, net capacity, whether a prior light water reactor existed at the same site, whether the location is in the north-east region of the United States, whether a cooling tower is used, whether the nuclear steam supply system was manufactured by Babcock-Wilcox, cumulative number of power plants constructed by each architectural engineer, whether there was a partial turnkey guarantee. Source: The data set is reported by Cox and Snell (1981) and Davison (2003). It is electronically available from R package SMPracticals (Davison, 2013), where it is labeled *nuclear*.

Rebellion Objects: 32 Romanian counties in 1907. Criterion: proportion of villages in which rebellious events took place in the Romanian peasant rebellion of 1907, labelled *spread*. Attributes: proportion of arable land devoted to wheat, proportion of rural population that is illiterate, strength of middle peasantry (measured by the proportion of land owned in units of 7 to 50 hectares), Gini coefficient of inequality of landownership,

population, region (Northern, South Central, Southwest, Eastern). Source: The data set was assembled by Chirot and Ragin (1975). Partial data set is reported by Fox (2008) and is electronically available from a website maintained by the author (Fox, 2015).

Recycle Objects: 31 Scottish local authorities. Criterion: weekly recyclate yield. Attributes: weekly recycling capacity, weekly residual capacity, number of principal materials collected, number of extended materials collected, frequency of recycling collection, frequency of residual collection, type of sort (comingled, curbside sort, dual service, single material). Source: The data were obtained by Baird et al. (2013) from Scottish local authorities. Partial data set is available electronically from an online repository maintained by Winner, where the data set is described as *recycling capacity, items collected and average yield for Scottish local authorities*.

Rent Objects: 2053 apartments in Munich, Germany. Criterion: rent per squaremeter in euros. Attributes: size, number of rooms, year of construction, whether the apartment is located at a good address, whether the apartment is located at the best address, whether the apartment has warm water, whether the apartment has central heating, whether the bathroom has tiles, whether there is special furniture in the bathroom, whether the apartment has an upmarket kitchen. Source: The data set is reported in Fahrmeir et al. (2010) and is electronically available from R package catdata (Schauberger and Tutz, 2014).

Salary Objects: 52 professors at a Midwestern college in the United States. Criterion: academic year salary. Attributes: sex, rank (assistant professor, associate professor, full professor), number of years in current rank, the highest degree earned (doctorate, masters), number of years since highest degree was earned. Source: The data set is reported by Weisberg (2005) and is electronically available from associated R package alr3 (Weisberg, 2011).

SAT Objects: 50 US states. Criterion: average total score on the SAT, 1994-95. Attributes: average expenditure per pupil, average pupil to teacher ratio, average salary of teachers, percentage of eligible students. Source: The data were collected by Guber (1999). The data set is electronically available from the R package faraway (Faraway, 2011).

Sperm Objects: 24 heterosexual couples. Criterion: mean sperm count per copulation. Attributes: age, height, and weight of each of the partners involved, volume of one male teste. Source: The data were collected by Baker and Bellis (1993). The data set is reported by Wood (2006) and is electronically available from associated R package gamair (Wood, 2012), where it is labeled *sperm.comp2*.

Tip Objects: 244 parties dining in a restaurant. Criterion: tip rate. Attributes: dollar amount of the bill, size of the party, sex of the bill payer, day of the week, time of the day, whether there were smokers in the party. Source: Data were recorded by a food server in a restaurant located in a suburban shopping mall in the United States during an interval of two and a half months in early 1990. The data set is reported in a collection of case studies for business statistics (Bryant and Smith, 1995). It is electronically available from R package reshape (Wickham, 2007).

Vote Objects: 159 counties in Georgia, USA. Criterion: proportion of uncounted votes in the 2000 presidential election. Attributes: type of voting equipment used (optical scan with central count, optical scan with precinct count, punch card, lever, paper), whether the county is in Atlanta, whether the county is urban or rural, proportion of African Americans, economic status (rich, middle, poor). Source: The data set was assembled by Meyer (2002). It is reported by Faraway (2005) and is electronically available from associated R package faraway (Faraway, 2011), where it is labeled *gavote*.

Waste Objects: 20 days of a laboratory experiment. Criterion: oxygen absorbed by dairy waste kept in suspension in water. Attributes: biological oxygen demand, chemical oxygen demand, total Kjedahl nitrogen, total solids, total volatile solids. Source: The data set is reported by Weisberg (2005) who notes that the data are from an experiment by Moore (1975). The data set is electronically available from R package alr3 (Weisberg, 2011) where it is labeled *dwaste*.

References

Joseph Adler. *R in a nutshell: A desktop quick reference*. "O'Reilly Media, Inc.", 2010.

Joseph Adler. *nutshell: Data for "R in a Nutshell"*, 2012. URL `http://CRAN.R-project.org/package=nutshell`. R package version 2.0.

Truett Allison and Domenic V. Cicchetti. Sleep in mammals: Ecological and constitutional correlates. *Science*, 194(4266):732–734, 1976.

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Jim Baird, Robin Curry, and Tim Reid. Development and application of a multiple linear regression model to consider the impact of weekly waste container capacity on the yield from kerbside recycling programmes in scotland. *Waste Management & Research*, 31(3): 306–314, 2013.

R. Robin Baker and Mark A. Bellis. Human sperm competition: Ejaculate adjustment by males and the function of masturbation. *Animal Behaviour*, 46(5):861–885, 1993.

W. John Braun. *MPV: Data Sets from Montgomery, Peck and Vining's Book*, 2012. URL `http://CRAN.R-project.org/package=MPV`. R package version 1.27.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Peter G. Bryant and Marlene A. Smith. *Practical Data Analysis: Case Studies in Business Statistics*. Richard D. Irwin Publishing, Homewood, IL, 1995.

Daniel Chirot and Charles Ragin. The market, tradition and peasant rebellion: The case of Romania in 1907. *American Sociological Review*, pages 428–444, 1975.

Singat Chu. Pricing the C's of diamond stones. *Journal of Statistics Education*, 9(2), 2001.

David R. Cox and E. Joyce Snell. *Applied Statistics: Principles and Examples*. Chapman and Hall, 1981.

Ann Crittenden. Vital dialogue is beginning between the rich and the poor. *The New York Times*, pages E–3, September 28 1975.

Yves Croissant. *Ecdat: Data sets for econometrics*, 2013. URL `http://CRAN.R-project.org/package=Ecdat`. R package version 0.2-2.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G. Goldstein. How good are simple heuristics? In Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group, editors, *Simple heuristics that make us smart*, pages 97–118. Oxford University Press, New York, 1999.

Anthony C. Davison. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. ISBN 0-521-77339-3.

Anthony C. Davison. *SMPracticals: Practicals for use with Davison (2003) Statistical Models*, 2013. URL `http://CRAN.R-project.org/package=SMPracticals`. R package version 1.4-2.

Stanley Dodson. Predicting crustacean zooplankton species richness. *Limnology and Oceanography*, 37(4):848–856, 1992.

David J. Edwards, Gary D. Holt, and Frank C. Harris. A comparative analysis between the multilayer perceptron "neural network" and multiple regression analysis for predicting construction plant maintenance costs. *Journal of Quality in Maintenance Engineering*, 6 (1):45–61, 2000.

Isaac Ehrlich. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy*, 81:521–565, 1973.

Phillip Ein-Dor and Jacob Feldmesser. Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, 30(4): 308–317, 1987.

Claus Ekstrom and Helle Sorensen. *isdals: Provides datasets for Introduction to Statistical Data Analysis for the Life Sciences*, 2014. URL `http://CRAN.R-project.org/package=isdals`. R package version 2.0-4.

Claus Thorn Ekstrom and Helle Sørensen. *Introduction to statistical data analysis for the life sciences*. CRC Press, 2010.

Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer Berlin Heidelberg, 2010. ISBN 9783642019388.

Ray C. Fair. A theory of extramarital affairs. *Journal of Political Economy*, 86:83–98, 1978.

Julian Faraway. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, 2005.

Julian Faraway. *faraway: Functions and datasets for books by Julian Faraway*, 2011. URL http://CRAN.R-project.org/package=faraway. R package version 1.0.5.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

John Fox. *Applied regression analysis and generalized linear models*. SAGE Publications, 2nd edition, 2008.

John Fox. Applied regression analysis and generalized linear models, second edition, data sets, 2015. http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/.

John Fox and Sanford Weisberg. *car: Data to accompany An R companion to applied regression 2rd edition*. Thousand Oaks CA, 2011a. URL http://CRAN.R-project.org/package=alr3. R package version 2.0.20.

John Fox and Sanford Weisberg. *An R companion to applied regression, Thousand Oaks, California*. Sage, Thousand Oaks CA, 2011b.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. *glmnet: Lasso and elastic-net regularized generalized linear models*, 2009. URL http://CRAN.R-project.org/package=glmnet.

Michael Friendly. *genridge: Generalized ridge trace plots for ridge regression*, 2012. URL http://CRAN.R-project.org/package=genridge. R package version 0.6-3.

Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

William H. Greene. Econometric analysis, 7th edition, links to data tables. http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

William H. Greene. *Econometric analysis*. Pearson Education India, 2003.

Deborah C. Guber. Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education*, 7(2), 1999.

David J. Hand, Fergus Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. *A handbook of small data sets*. Chapman & Hall/CRC, London, UK, 1994.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2. edition, 2009.

Clifford G. Hildreth and John Y. Lu. Demand relations with auto-correlated disturbances. Technical Bulletin 276, Michigan State University, Agricultural Experiment Station, Department of Agricultural Economics, East Lansing, Michigan, November 1960.

David C. Hoaglin and Paul F. Velleman. A critical look at some analyses of major league baseball salaries. *American Statistician*, 49(4266):277–285, 1995.

David Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. Wiley, 2000.

David R. Jarvis. Nitrogen levels in long bones from coffin burials interred for periods of 26–90 years. *Forensic Science International*, 85(3):199–208, 1997.

Michael P. Johnson and Peter H. Raven. Species number and endemism: The Galápagos archipelago revisited. *Science*, 179(4076):893–895, 1973.

Koteswara Rao Kadiyala. Testing for the independence of regression disturbances. *Econometrica*, 38(1):97–117, 1970.

Sanford Labovitz. The assignment of numbers to rank order categories. *The American Sociological Review*, 35:515–524, 1970.

Samuel Leinhardt and Stanley S. Wasserman. Exploratory data analysis: An introduction to selected methods. *Sociological methodology*, 10:311–365, 1979.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Robin H. Lock. 1993 New car data. *Journal of Statistics Education*, 1(1), 1993.

John Maindonald and W. John Braun. *Data Analysis and Graphics Using R: An Example-Based Approach*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

John Maindonald and W. John Braun. *DAAG: Data Analysis And Graphics data and functions*, 2013. URL `http://CRAN.R-project.org/package=DAAG`. R package version 1.16.

Gary C. McDonald and Richard C. Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463–482, 1973.

Lauren McIntyre. Using cigarette data for an introduction to multiple regression. *Journal of Statistics Education*, 2(1), 1994.

William Mendenhall and Terry Sincich. *Statistics for Engineering and the Sciences*. Dellen Publishing, New York, 3rd edition, 1992.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6-4.

Mary C. Meyer. Uncounted votes: Does voting equipment matter? *Chance*, 15(4):33–38, 2002.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 3rd edition, 2001.

J. A. Moore. *Total biomedical oxygen demand of animal manures*. PhD thesis, University of Minnesota, 1975.

Felicia B. Morton. Charting a school's course. *Chicago*, pages 86–95, 1995.

Yuzo Murayama. Information and emigrants: Interprefectural differences of japanese emigration to the pacific northwest, 1880–1915. *The Journal of Economic History*, 51(1): 125–147, 1991.

Raymond H. Myers. *Classical and modern regression with applications*. Duxbury advanced series in statistics and decision sciences. PWS-KENT, second edition, 1990.

Keith W. Penrose, A. G. Nelson, and A. G. Fisher. Generalized body composition prediction equation for men using simple measurement techniques. *Medicine & Science in Sports & Exercise*, 17(2):189, 1985.

John A. Rice. *Mathematical Statistics and Data Analysis*. Wadsworth Inc., Belmont, CA, 1995.

Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, and David Firth. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, 2013. URL `http://CRAN.R-project.org/package=MASS`. R package version 7.3-28.

Dennis Rodkin. 10 Keys for creating top high schools. *Chicago*, pages 78–85, 1995.

Gunther Schauberger and Gerhard Tutz. *catdata: Categorical Data*, 2014. URL `http://CRAN.R-project.org/package=catdata`. R package version 1.2.1.

J. John Sepkoski and Michael A. Rex. Distribution of freshwater mussels: coastal rivers as biogeographic islands. *Systematic Biology*, 23(2):165–188, 1974.

Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer, 1996.

Jeffrey S. Simonoff. Online data archive. `http://people.stern.nyu.edu/jsimonof/SmoothMeth/`, 2015.

Gordon K. Smyth. Australasian Data and Story Library (OzDASL), 2011. URL `http://www.statsci.org/data`.

Robert R. Sokal and F. James Rohlf. *Biometry: The principles and practice of statistics in biological research*. W. H. Freeman and Company, San Francisco, 2nd edition, 1981.

William L. Stanley and Michael Douglas Miller. Measuring technological change in jet fighter aircraft. Technical Report R-2249-AF, RAND Corporation, Santa Monica, CA, 1979.

StatLib: Data, software and news from the statistics community, 2013. URL `http://lib.stat.cmu.edu/`.

Ira B. Tager, Sscott T. Weiss, Bernard Rosner, and Frank E. Speizer. Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, 110(1):15–26, 1979. ISSN 0002-9262.

Richard D. Telford and Ross B. Cunningham. Sex, sport, and body-size dependency of hematology in highly trained athletes. *Medicine and Science in Sports and Exercise*, 23: 788–794, 1991.

William Tucker. Where do the homeless come from? *National Review*, pages 34–44, 1987.

Read D. Tuddenham and Margaret M. Snyder. Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in child development*, 1(2):183, 1954.

B. L. Turner, Robert Q Hanham, and Anthony V Portararo. Population pressure and agricultural intensity. *Annals of the Association of American Geographers*, 67(3):384–396, 1977.

United States Navy. *Procedures and analysis for staffing standards development: Data/Regression analysis handbook*. Navy Manpower and Material Analysis Center, San Diego, CA, 1979.

Walter Vandaele. Participation in illegitimate activities: Ehrlich revisited. In A. Blumstein, J. Cohen, and D. Nagin, editors, *Deterrence and Incapacitation*, pages 270–335. National Academy of Sciences, Washington DC, 1978.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

Sanford Weisberg. *Applied Linear Regression*. Wiley, Hoboken NJ, 3rd edition, 2005.

Sanford Weisberg. *alr3: Data to accompany Applied Linear Regression 3rd edition*, 2011. URL `http://CRAN.R-project.org/package=alr3`. R package version 2.0.5.

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007. URL `http://www.jstatsoft.org/v21/i12/paper`.

Larry Winner. Miscellaneous datasets. `http://www.stat.ufl.edu/~winner/datasets.html`.

Simon Wood. *gamair: Data for "GAMs: An Introduction with R"*, 2012. URL `http://CRAN.R-project.org/package=gamair`. R package version 0.0-8.

Simon N. Wood. *Generalized additive models: An introduction with R*. Chapman & Hall/CRC, London, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.