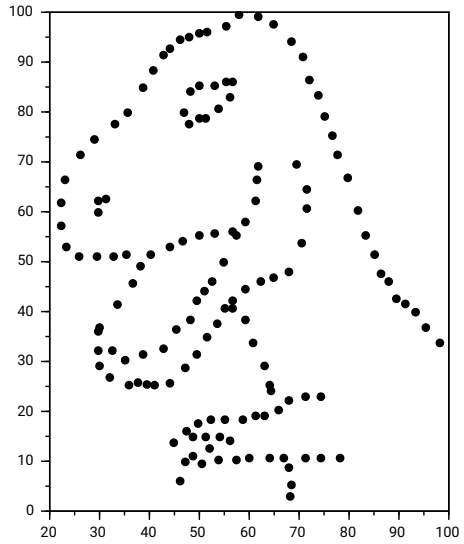
Kap. 1 Anhang 1: Der Datasaurus und seine Verwandten

Alberto Cairo, aus Spanien stammender und in den USA lehrender Professor für *Informations-Design* – dazu zählt insbesondere auch die Visualisierung statistischer Daten –, hat 2016 ein anschaulich sehr schönes Beispiel einer *linear unkorrelierten* Punktwolke publiziert, die er *Datasaurus* nannte:



Die wichtigsten Kenngrößen dieses aus 142 Koordinatenpaaren $[x_i, y_i]$ bestehenden Datensatzes:

$$\bar{x} = 54.26..., \ \bar{y} = 47.83..., \ s_x = 16.76..., \ s_y = 26.93..., \ r_{xy} = -0.06...$$

Die Botschaft ist: Ein paar Kenngrößen allein, so wichtig sie manchmal sind, kennzeichnen einen Datensatz nur sehr unvollkommen; wann immer möglich, sollte man eine *Visualisierung* anstreben. Obwohl praktisch linear unkorreliert, da $r_{xy} \approx 0$, sind die 142 Datenpunkte *nicht unabhängig* und stehen auf ganz spezielle Weise in einem nichtlinearen Zusammenhang.

Zwei kanadische Statistik-Experten, Justin Matejka und George Fitzmaurice, haben diesem Datensatz *ein Dutzend* weitere an die Seite gestellt, die alle jeweils aus 142 Datenpaaren bestehen und *bis zur zweiten Nachkommastelle mit denen des Datasaurus übereinstimmende* Kenngrößen $\bar{x}, \bar{y}, s_x, s_y, r_{xy}$ besitzen.

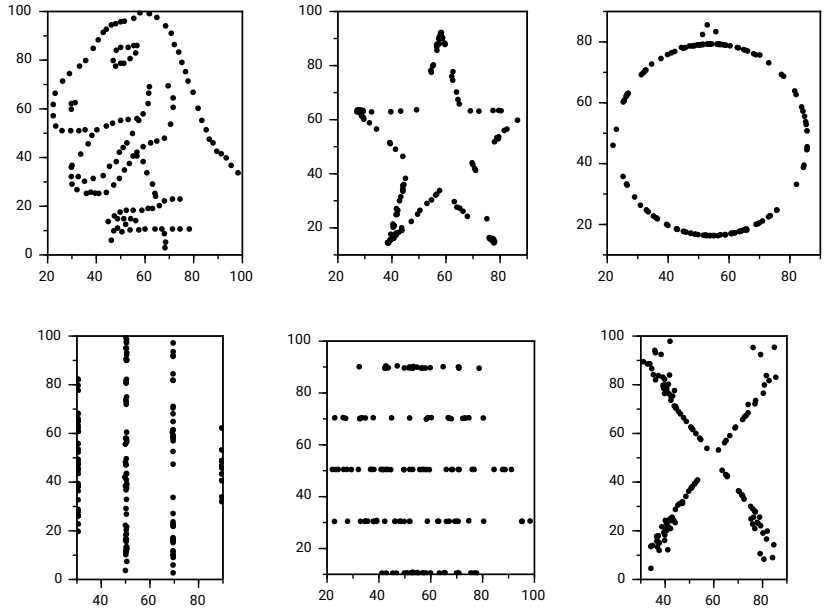
Die Datensätze wurden durch einen aufwendigen numerischen Optimierungsprozess erzeugt:

Per iterativer *zufälliger* Abänderung der x_i und y_i unter Beibehaltung der Nebenbedingungen $\bar{x} = 54.26..., \ \bar{y} = 47.83..., \ s_x = 16.76..., \ s_y = 26.93..., \ r_{xy} = -0.06...$ wurde versucht, sich an das jeweils vorgegebene Verteilungsmuster anzunähern.

Damit die schrittweise Optimierung nicht in einem nur lokalen Neben-Optimum hängenbleibt, wurde eine *Simulated-Annealing*-Komponente in die Iterationen eingebaut. Pro Zielfigur wurden ca. 200 000 Iterationsschritte benötigt!

Die so erzielten sehr erstaunlichen zwölf Datensätze von je 142 Punkten nannten die Autoren **The Datasaurus Dozen**.

Webseite: https://www.autodesk.com/research/publications/same-stats-different-graphs#



Hier wurde mittels SCILAB (Befehle subplot(m,n,k) und scatter) eine Auswahl von fünf der zwölf Datensätze dem Datasaurus, dem Ausgangs-Datensatz, gegenübergestellt. Alle linear unkorreliert, aber doch offensichtlich auf jeweils ganz unterschiedliche Weise (nichtlinear) abhängig.

Natürlich sind Kenngrößen *im Normalfall* dennoch aussagefähig, insbesondere dann, wenn man *Vorwissen* über die Verteilung der x_i und y_i hat. Aber generell ist die Visualisierung der Daten meist von großem Erkenntnis-Wert. Wie ja auch die *Lorenzkurve weit aussagefähiger* ist hinsichtlich der Einkommens- und Vermögensverteilung als der oft genannte *Gini-Koeffizient*.

Auf "The Datasaurus Dozen" wurde ich aufmerksam durch einen Hinweis in dem exzellenten Buch

The Art of Statistics - Learning from Data
von Sir **David John Spiegelhalter**
Pelican Books 2020

Spiegelhalter ist Lehrstuhlinhaber für Statistik an der Universität Cambridge. "Spiegelhalter is one of the most cited and influential researchers in his field, and was elected as President of the Royal Statistical Society for 2017-18."