Graphing and Communicating Compositional Data in High Dimensions

H.-F. ULBRICH

Global Drug Discovery Statistics – Bayer HealthCare, Germany, hannesfriedrich.ulbrich@bayer.com

Visualization of data becomes more challenging as the dimensionality of the data increases, impacting not only the display of the data itself but also the modeling results.

This paper discusses common visualization techniques for compositional data. None of them seem to be well suited for changes in compositions that depend on either a metric covariate or a factor. The **clr**-deviation chart as a chart with a factor or covariate as abscissa and all centered log ratio-transformed component values superimposed on the ordinate axis is then introduced jointly with the **clr**-component chart. The **clr**-deviation chart takes advantage of the sum-equals-zero property of **clr**-transformed compositional data. It has some theoretical and practical advantages over alternatives and one major disadvantage – an arbitrarily scaled ordinate axis; its properties are discussed.

The usefulness of the methods are illustrated using an example analyzing the changes of proportions of the different diseases treated by hospitalization over a period of 13 years in Germany.

Introduction

Visualizing data is always a challenging task. Data evaluation can be efficiently supported by visualizing the raw data for consistency checks and exploration of inherent patterns. Communication of statistical models, parameters and predictions shall be accompanied by appropriate graphical representation of analysis results.

Graphing data becomes more challenging as the dimensionality increases. Compositional data as multi-component data of a constant sum are multi-dimensional by nature. The constant sum property for D components restricts the data to a simplex $S^D = \{x \in \mathbb{R}^D \mid x_i > 0,\ x_1 + \ldots + x_D = const\}$ within the D-dimensional Euclidean space $\mathbb{R}^D$. Without loss of generality let's set the constant to 1 ($= 100\,\%$). Any compositional vector $\vec{x} = (x_1, \ldots, x_D)' \in S^D$ is a point of the simplex.

Ternary diagrams are a suitable means for exploratory data analysis of compositional data (Pawlowsky-Glahn et al., 2007). Here the equilateral triangle is equivalent to the simplex S^3 accommodating three components such that each vertex of the triangle represents a point where either of the components takes the maximum value of 1, the opposed edge represents all points where that particular component has the minimum value of 0. Since the ternary diagram is rotationally symmetric by 120° around the barycenter as well as symmetric around any angle bisector it is fair with respect to the interchangeability of the components, i. e., with regard to the visual quality ternary diagrams are permutation invariant.

Although less common one could use the regular tetrahedron in a similar and fair (symmetric between components) fashion as before as a representation for the $S^4 \subset \mathbb{R}^4$ using a tool for dynamic graphs.

For more than four components visualization of compositional data must rely on other techniques. Biplots tailored for compositional data were introduced by Aitchison and Greenacre (2002), they become better known as log-ratio biplots (Greenacre, 2010, chapter 7). They are a projection of both the compositions of a sample and all the components on a plane and therefore a means of dimension reduction.

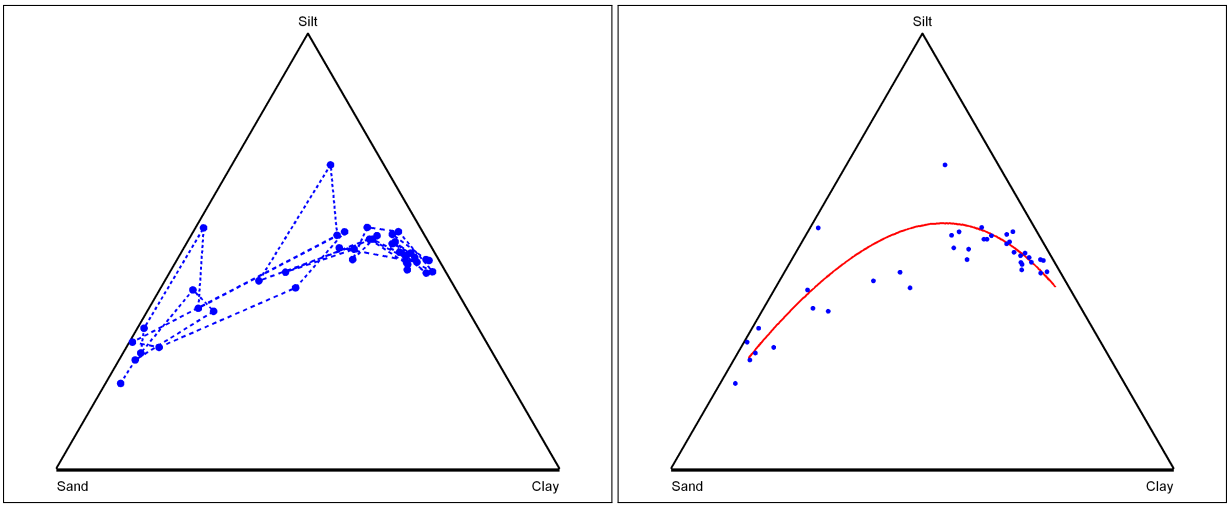
Two different techniques of dimensionality reduction have already been defined in Aitchison (1986, 2003, sections 2.5 and 2.6): subcomposition and amalgamation. Analyzing subcompositions means giving attention to the relative magnitudes of a subset of components (by ignoring the remaining). Amalgamation is reducing the number of components by adding up two or more into an amalgamated component.

Subcompositions and amalgamated compositions of three components might be visualized by aforementioned ternary diagrams leaving out some parts of the original information.

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

Visualizing compositions along with an influential factor

Graphing gets more complex if one wants to visualize changes in compositions with regard to changes in either a metric covariate or different levels of a factor. Fig. 1 shows changes in sand, silt, clay compositions of sediment samples at different water depths in an Arctic lake (Aitchison, 2003, data set 5, p. 359).

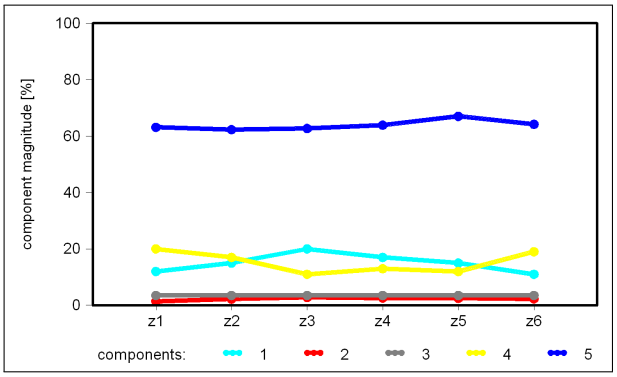


(a) (b)

Fig. 1: Changes in sand, silt, clay compositions of 39 sediment samples at different water depth in an Arctic lake
(a) measured data (connected in order of increasing water depth)
(b) measured data and regression curve, log-depth as regressor

Panel (b) of fig. 1 shows the expected compositional triplets (sand,silt,clay) as function of the logarithm of water depth; from the graph it is neither obvious that the sand component is decreasing with increasing water depth nor how good the regression curve fits the data.

factor level	components [%]				
	#1	#2	#3	#4	#5
z1	12.0	1.4	3.5	20.0	63.1
z2	15.0	2.2	3.5	17.0	62.3
z3	20.0	2.8	3.5	11.0	62.7
z4	17.0	2.6	3.5	13.0	63.9
z5	15.0	2.5	3.5	12.0	67.0
z6	11.0	2.3	3.5	19.0	64.2

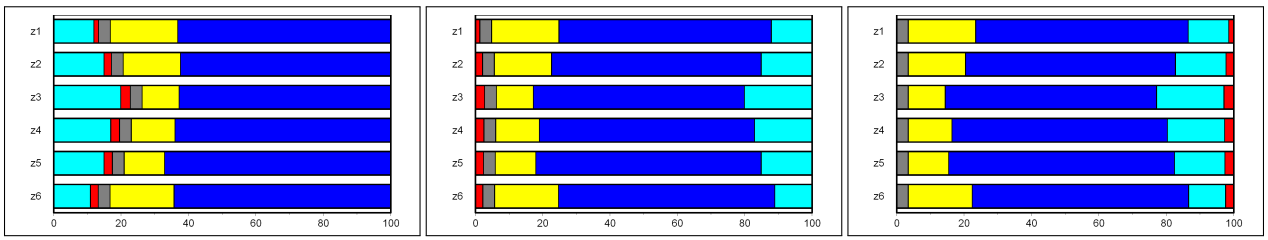


(a) (b)

Fig. 2: Compositions up to 100 %: (a) artificial example data, (b) superimposed line chart

For compositions of dimensions higher than 3 and to put more focus on the direction of increasing values of the covariate line charts are commonly used. For visualization of a covariate's influence on compositions, these line charts have the covariate on abscissa axis and the components' values on ordinate axis, the latter either in superimposed or stacked manner. Superimposed here means all components are shown as magnitude above zero, i. e., to leave out from the diagram the constant sum property. Fig. 2 gives an artificial example of 5-component compositional data along with a related superimposed line chart.

Stacked line charts as well as stacked bar charts try to visualize components with regard to their sum and therefore seem to be preferable for compositional data. Bar charts — either vertically or horizontally stacked — suit better to levels of a factor or discrete fixed in advance steps of a metric variable like time measured in years.



(a) (b) (c)

Fig. 3: Compositions up to 100 % (data as in fig. 2) — different orders of the components
(a) 1–2–3–4–5, (b) 2–3–4–5–1, (c) 3–4–5–1–2

As can be seen in fig. 3, the impression generated by stacked bar charts depends heavily on the order of the components in the diagram. The same holds true with stacked line charts. Changes in components anchored either at 0 or max (100 % here) can better be judged visually than any other component in-between. Neither stacked bar charts nor stacked line charts are fair with respect to the interchangeability of the components.

Even more, with an increasing number of dimensions, a component that is low in magnitude become rarely visible, although its original value might 'double' (as component # 2 of the example: 1.4 at level 'z1', 2.8 at level 'z3') and therefore has quite an impact on the changes of the complete composition itself.

Compositional deviation charts shall overcome these disadvantages.

The compositional or clr-deviation chart

Aitchison (1986) introduced two different classes of bijective transformations between the (topologically open) simplex S^D and a (D-1)-dimensional Euclidean space, the additive and the centered log-ratio transformations. The class of **alr**-transformations consists of D different transformations depending on which of the components $x_1, \ldots, x_D$ will be used as reference component. Any of the **alr**-transformations maps the simplex S^D, the components of each composition, into coordinates of the $\mathbb{R}^{D-1}$. This transformation is not isometric with regard to the Aitchison metric on the simplex (Pawlowsky-Glahn et al., 2007). The class of **clr**-transformations consists of exactly one member mapping S^D isometrically to a (D-1)-dimensional hyperplane through the origin of $\mathbb{R}^D$ (with the geometric mean $\mathrm{g}(\vec{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ being a scalar):

$$\mathbf{clr}(\vec{x}) = \ln\left(\frac{\vec{x}}{\mathrm{g}(\vec{x})}\right) = \frac{1}{D} \cdot \begin{pmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & D-1 \end{pmatrix} \cdot \ln(\vec{x})$$

Performing **clr**-transformation on a D-composition results in coordinates within $\mathbb{R}^D$ such that their sum equals zero. Remarkable properties of **clr**-transformation therefore are: (a) the number of coordinates equals the number of components, (b) each component relates directly one-to-one to one coordinate, (c) it is symmetrical in the components, (d) it is an isometric transformation, and (e) each logarithmized component value gets standardized by subtracting from it the logarithm of the geometric mean $\mathrm{g}(\vec{x})$ of all the components of that particular composition.

The zero-sum property of the **clr**-transformed components can be expressed by $\mathbf{1}'_D \cdot \mathbf{clr}(\vec{x}) = 0$, the equation defining that aforementioned (D-1)-dimensional hyperplane through the origin of $\mathbb{R}^D$.

More than a decade later Egozcue et al. (2003) introduced a third class of bijective transformations between S^D and a (D-1)-dimensional Euclidean space called isometric log-ratio transformation. **ilr**-transformations are distance-preserving. Each orthonormal rotation of an **ilr**-transformation belongs to the class of **ilr**-transformations.

As is explained in Pawlowsky-Glahn et al. (2007, sect. 4.4) there is always a (D-1)×D-dimensional matrix Ψ such that both $\Psi \cdot \Psi' = \mathbf{I}_{D-1}$ and $\Psi' \cdot \Psi = \mathbf{I}_D - \frac{1}{D} \cdot \mathbf{1}'_D \cdot \mathbf{1}_D$ are satisfied. This matrix Ψ

can be used for a rotation between the (D-1)-dimensional **clr**-hyperplane and the (D-1)-dimensional Euclidean space $\mathbb{R}^{D-1}$ forming an isometric link between **clr**- and the class of **ilr**-transformations:

$$\mathbf{ilr}\left(\vec{x}\right) = \Psi \cdot \mathbf{clr}\left(\vec{x}\right) \qquad \text{and} \qquad \mathbf{clr}\left(\vec{x}\right) = \Psi' \cdot \mathbf{ilr}\left(\vec{x}\right)$$

Log-ratio coordinates and coefficients of random composition are real random variables ranging freely form $-\infty$ to ∞. Therefore, it is common to analyze them with standard multivariate procedures. For **ilr**-coordinates these methods can be used straightaway (Pawlowsky-Glahn and Egozcue, 2006). Furthermore, data in **ilr**-coordinates as well as residuals of **ilr**-based models can be visualized by orthogonal projection of the (D-1)-dimensional Euclidean space onto lower dimensional subspaces. However, interpretation of these graphs in terms of the compositions' components is difficult because coordinates refer to 'mixtures' of the components.

Orthogonally projecting **clr**-coordinates instead gives similar graphs; due to the direct one-to-one relationship between component and coordinate these graphs are easier to interpret in terms of the original components. Line charts showing one **clr**-coordinate on ordinate and either different levels of a factor or a metric covariate on abscissa axis are of particular interest here.

These **clr**-coordinate charts can be used in the usual way for visualizing components of raw data over a predictor — with or without a regression line . Since they are linear projections of an Euclidean space they can also be used for plotting residuals against the predictor as in common multivariate regression.

There are D component specific **clr**-coordinate charts. Unlike in the common multivariate situation where each dimension is a variable likely to be on a different scale (and therefore quite often dimension-wise standardized by $\frac{x_i - \overline{x}}{s}$ for making them comparable), all **clr**-coordinate charts are already composition-wise standardized by $-\ln(\mathrm{g}\left(\vec{x}\right))$. Each of the **clr**-coordinate charts addresses the component's deviation from the geometric mean $\mathrm{g}\left(\vec{x}\right)$ by showing the coordinate's deviation from 0.

A **clr**-deviation chart (or compositional deviation chart) is the superposition of all related **clr**-coordinate charts. It shows jointly the deviation of any of the **clr**-coordinates of compositions against levels of a factor or an interval-scaled covariate as (potential) predictor (fig. 4).

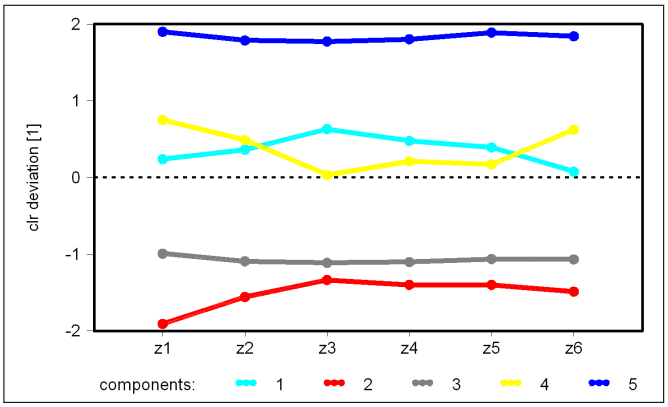


Fig. 4: **clr**-deviation chart (data from fig. 2, connecting lines for improved readability)

The **clr**-deviation chart it is symmetrical in the components and fair, any changes of the order of the components does not change the **clr**-deviation chart at all.

Because of the sum-equals-zero property **clr**-deviation charts do not only show changes in certain **clr**-coordinates and therefore their related components but also the components 'compensating' for these changes. Because of the isometric relation between **clr**- and **ilr**-coordinates compositional deviation charts show either data as they will be analyzed in coordinate space or modeling results (regressions) of these analyses. Panels of **clr**-deviation charts can be used for visual comparison of different compositional time series as well as group specific regression results of compositional panel or repeated (over time) measurement data, amongst others.

Two properties might become a little more difficult for communicating compositional data analysis by **clr**-deviation charts. Firstly, the scale of the ordinate axis seems to be arbitrary, and secondly, **clr**-deviation charts present data not on the original scale of the component values. Graphs based

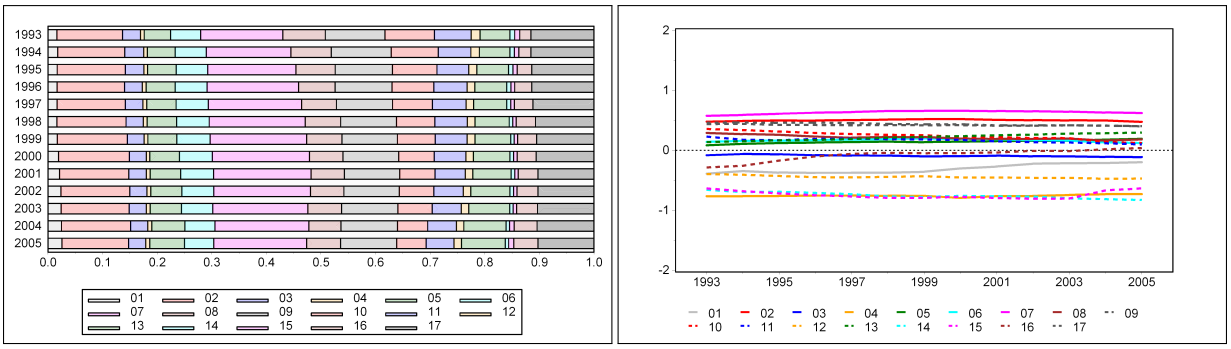
on transformed scales are quite often considered to be challenging to the not so experienced viewer, compositional deviation plots share this obstacle with ternary plots and others. Analyzing compositional data properly requires a minimum understanding of Aitchison geometry and the staying-in-the-simplex approach (Aitchison and Egozcue, 2005, and the therein reference on p. 831 to for years discussions in Mathematical Geology).

Example

Since the early 1990s Germany has had a yearly requirement for all infirmaries hospital statistics in three parts: basis data (location, staff and equipment, number of cases), cost data, and diagnosis data. The latter consists of all hospital stays (cases) concluded in that particular year, each case is recorded with some basic demographics and its main diagnosis. The diagnoses get coded by the hospitals themselves. Up to 1999 the main diagnosis had been coded according to the ICD-9 coding scheme, after that the ICD-10 coding scheme was to be used. According to both coding schemes diseases are categorized into chapters (e. g., 'injury and poisoning') and subchapters (e. g., 'fractures', 'open wounds', 'burns', etc.). ICD-9 in its most general form is a numeric code of 3-digit numbers between 001–'cholera' and 999–'complications of medical care, not elsewhere classified'; overall about 820 codes (with some omissions within the number range). It had been replaced by an alphanumeric one of one leading letter followed by two digits; mapping ICD-10 codes onto ICD-9 ones is said to be possible and supported by an algorithm. (For data non-disclosure reasons according to German legal regulations only the 3-digit disease codes have been made available for evaluation.)

Analyzing diagnosis data of the German hospital statistics from an econometric perspective we became interested in changes of proportions of the different diseases treated by hospitalization over a period of 13 years (1993–2005 inclusive). Disease data (the main diagnosis per case only) is classified and coded; it is compositional in such that it is closed by the yearly number of concluded hospital stays: at a particular hospital, in all hospitals of an administrative region, to name a few.

Seven (of the 16) federal states have been chosen (including one of the 3 city states) for the analysis. Since the number of cases without a given diagnosis is negligible it is not considered here.

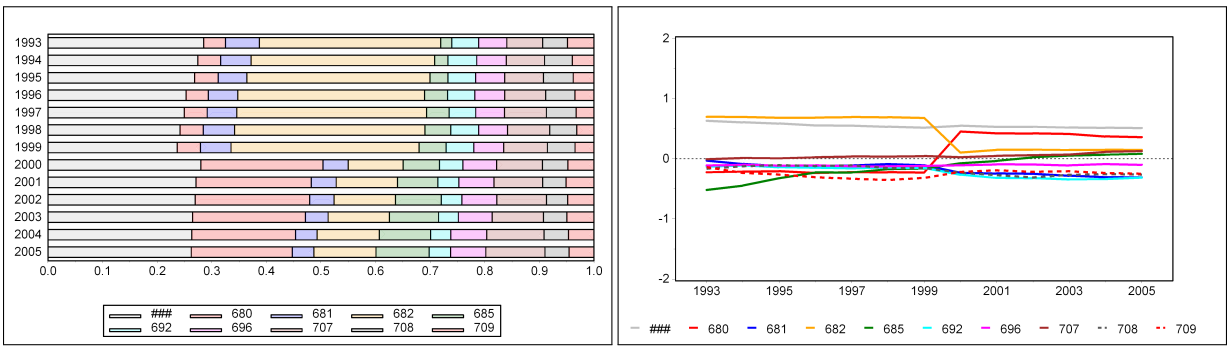


(a) (b)

Fig. 5: composition of the number of diagnoses in 17 chapter of ICD-9
(a) stacked bar chart, (b) clr-deviation chart

The most general overview of all diagnosis data can be given by dividing all the diagnoses coded in German hospitals into the 17 ICD-9 chapters (from 01–'infectious and parasitic diseases' to 17–'injury and poisoning'). Any chapter as component of the composition here can be seen as amalgamation of all diseases coded belonging to that chapter. Results are shown in fig. 5 indicating that over that 13 year period only marginal changes in diagnosis-chapter composition took place in German hospitals, as can be seen in both panels.

Since the diagnoses of particular chapters and subchapters are of interest on their own, any of these can be analyzed as a subcomposition. Chapter 12–'diseases of the skin and subcutaneous tissue' covers the disease codes 680–709 (26 different codes, 687–689 and 699 not being valid code numbers). Codes 680, 681, 682, 685, 692, 696, 707, 708, 709 are the most often registered ones; all others have been amalgamated into one component (marked by '###').

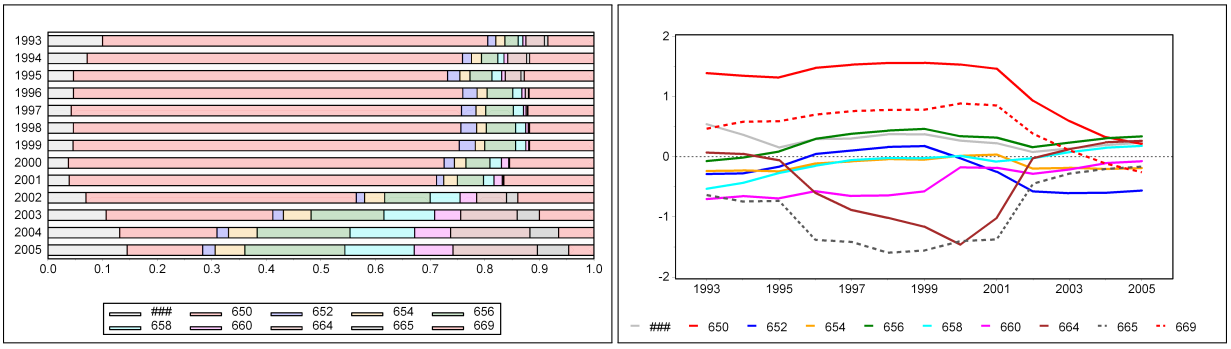


(a) (b)

Fig. 6: ICD-9 chapter 12–'diseases of the skin and subcutaneous tissue' (###: amalgamated others)
(a) stacked bar chart, (b) clr-deviation chart

Both panels of fig. 6 show discontinuities in the yearly changes in composition, panel (b) highlights the fact that the decrease in diagnosis code 682 is mostly corrected by an increase in diagnosis code 680, although not fully symmetric. Some dermatological patients of the year 2000 and later seem to be differently ill as compared to the previous period. This 'effect' would be of particular interest for both epidemiologists as well as economists since there is no other remarkable change visible over the whole period and that change might have economic consequences as well. Since the 'effect' coincides fully with the change in the disease coding system by the end of 1999 checking the ICD-10-to-ICD-9 mapping algorithm seems recommended in the first instance. It reveals that there is a shift in what is covered by these particular ICD-9 codes and their ICD-10 counterparts, where code 680 covers one particular sub-disease (indication, symptom) less than its counterpart; this sub-disease is now to be covered by the ICD-10 counterpart of code 682. The 'effect' therefore turns out to be caused by the change in the coding system and cannot be resolved by the mapping algorithm.

More interesting from both the view of an economist and of an epidemiologist is the data related to hospitalized woman because of pregnancy and labor. Two subchapters of chapter 11–'complications of pregnancy, childbirth, and the puerperium' are devoted to the course of labor and delivery: codes 650–659 describe different indications of 'normal delivery, and other indications for care in pregnancy, labor, and delivery' whereas codes 660–669 are to be used for 'complications occurring mainly in the course of labor and delivery'–both number ranges are valid codes without any omissions, it is 20 different codes in total. (No code mapping inconsistencies have been detected here.)



(a) (b)

Fig. 7: ICD-9 subchapters of labor, delivery and complications with them (###: amalgamated others)
(a) stacked bar chart, (b) clr-deviation chart

As can be seen in the **clr**-deviation chart of fig. 7 there are two diagnoses becoming steadily less used during the 1990s: code 664 'trauma to perineum and vulva during delivery' and code 665 'other obstetrical trauma' in an otherwise rather unchanged composition over time. But starting 2001 remarkable changes in magnitude of some components (diagnose codes) occur with a considerable shift in trend. First and overwhelmingly visible just in the stacked bar chart the proportion of code 650–'normal delivery' drops from about 65 % in 2001 to less than 17 % in 2005. As can be seen in the

clr-deviation chart a similar drop occurred with code 669–'other indications for care or intervention related to labor'–an indication likely to be not too far from 'normal delivery'. (The share of deliveries in hospitals among all deliveries remained quite constant over the whole period.) The decrease in codes 650 and 669 is mainly compensated by an increase in before mentioned codes 664 and 665 while the slow decrease in code 652–'malposition and malpresentation of fetus' since the mid-1990s comes to an halt.

Surprisingly, as of now this change away from 'normal delivery' seems not to be yet well reflected in neither epidemiologic, health care management nor economic literature, although it should be of epidemiological and health policy concern. (For a first discussion of the changes until the end of 2003 see Heller and Schmidt, 2005.)

With the Statutory Health Insurance Reform Act 2000 the German government started to introduce a new case-based hospital funding system by adapting an internationally used diagnosis related groups (DRG) system. DRGs are meant to reflect the diagnosed disease(s) and the costs for treating them at hospital. Development of the German refined DRG System (G-DRG) started in June 2000; providing data by the German hospitals started on voluntary basis in 2003 and became mandatory in 2004 with the intention to accumulate data for calibrating the reimbursement rates for DRGs by keeping the overall hospital costs constant for the calibration period.

Further econometric analyses of the data at hand, e. g., differences between hospitals of different regions, is needed for better understanding of reasons and impact of the massive changes in diagnosis codes immediately before and during the first DRG-calibration period.

Conclusion

The **clr**-deviation chart has proven itself to be very useful. As a superposition of all **clr**-coordinate charts it is fair since it is symmetrical in the components of a composition. It can be used for data checking, presentation of raw data as well as model predictions.

clr-coordinate charts themselves are more likely to be helpful for component related residual analyses. Both charts rely on orthogonal projections and the isometric relationship (rotation) between **ilr**- and **clr**-coordinates.

Although being visualizations of transformed data both diagrams are appropriate means for communicating compositional data and model predictions. They show (projections of) the data as it is, or will be, analyzed.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data (Reprint with additional material)*. Caldwell, NJ: The Blackburn Press.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: Where are we and where should we be heading. *Mathematical Geology 37*, 829–850.

Aitchison, J. and M. Greenacre (2002). Biplots of compositional data. *Journal of the Royal Statistical Society C (Applied Statistics) 51*, 375–392.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*, 279–300.

Greenacre, M. (2010). *Biplots in Practice*. BBVA Foundation Manuals. Barcelona: Fundación BBVA.

Heller, G. and S. Schmidt (2005). Wodurch ist die Veränderung der geburtshilflichen Diagnosehäufigkeiten in der Krankenhausdiagnosestatistik zu erklären? In J. Klauber, B.-P. Robra, and H. Schellschmidt (Eds.), *Krankenhaus-Report 2005. Schwerpunkt Wege zur Integration*, pp. 297–300. Stuttgart, New York: Schattauer.

Pawlowsky-Glahn, V. and J. J. Egozcue (2006). Compositional data and their analysis: an introduction. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (Eds.), *Compositional Data Analysis in the Geosciences: From Theory to Praxis*, Volume 264 of *Geological Society Special Publication*, pp. 1–10. London: Geological Society.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2007). Lecture Notes on Compositional Data Analysis. `http://dugi-doc.udg.edu/bitstream/10256/297/1/CoDa-book.pdf` (13. 06. 2008).

Ulbrich, H.-F. (2010). Höherdimensionale Kompositionsdaten – Gedanken zur grafischen Darstellung und Analyse. Statistische Diskussionsbeiträge 43, Universität Potsdam, Potsdam.