Visualizing Compositional Data on the Lexis Surface*

Jonas Schöley† and Frans Willekens‡

schoeley@demogr.mpg.de · willekens@demogr.mpg.de

July 16, 2015

ABSTRACT. The analysis of compositional data is a topic inherent to demography. In recent times, due to a growing catalogue of detailed population data, it became feasible to consider populations not only structured by time, age or sex, but by any number of interesting criteria. This "inflation" of data dimensions produces challenges in visualizing the data. To aid the understanding of age-structured timelines of compositions we seek to extend the Lexis surface plot from 1-dimensional continuous data to multidimensional compositional data. We apply different strategies for visualizing compositional data on the Lexis surface to French death counts given by cause and compare the results for compliance with multiple desired criteria.

KEYWORDS. *visualization, colour schemes, compositional data, Lexis surface, causes of death, mortality*

From the display of population numbers by shading map regions, the graphical representation of population dynamics on a grid of year-age-cohort parallels, and the widely recognized population pyramid to today's heatmaps of mortality surfaces: Demography has always had a close relationship with information visualization.[1] The visual display helps making sense of the data at hand which in demography, for the most part, are *counts, rates and proportions.* Visualisation methods that are currently used in demography present counts or rates. This paper is about *compositional data*, represented by proportions, i. e. shares of a whole. Examples for this data type are proportions within a population (e. g. age composition, distribution by occupation, region of residence or level of education), proportions of events (e. g. deaths by cause),

---

[1]One of the first choropleth maps of population densities can be found in d'Angeville (1836). The year-age-cohort grid is commonly attributed to Lexis (1875) but for a full account of the inventors of the "Lexis"-diagram cp. Vandeschrick (2001). Population pyramids were first published in Walker (1874). In 1987 Vaupel, Gambill, and Yashin (1987) discussed the use of heatmaps in a demographic context.

proportions of durations (e. g. life expectancy by health status), and proportions within a total rate (e. g. death rate by cause of death).

While rates and counts provide a *single value* for each point on the Lexis-surface a *vector of values* is needed in the case of compositional data. Classical solutions for the visualization of population dynamics, such as contour maps or a one-dimensional heatmap of continuous data on the Lexis-surface, do not work for compositional data. On the other hand graphs specifically designed for compositional data like the ternary coordinates (Aitchison 1986:5 ff) or the biplot (Gabriel 1971; Aitchison 2002) do not address the basic demographic dimensions period, age and cohort and therefore are unsuited to show corresponding effects in a single display.

This paper aims to extend the visual repertoire of demography by introducing and discussing different techniques of graphing compositional data on the Lexis surface. Hereby we hope to facilitate the exploratory analysis of compositional data and the communication of research results in graphical form. To demonstrate the techniques we use publicly available data on age-specific death counts by cause of death in France from 1925 to 1999 (Vallin and Meslé 2014). Four visualizations are discussed in this paper. The first is the three-variable balance scheme or *ternary balance scheme*.[2] It is a technique to display three attributes in a single point. Each attribute is mapped to a primary colour and the mixture of three colours shows the composition of attributes in a population. The second visualization is the *qualitative sequential scheme*. In that scheme, a qualitative or categorical variable (e. g. cause of death) is represented by a colour and the quantitative variable (e. g. number of deaths due to that cause) is represented by sequences of lightness steps within each colour. The third visualization, the *age-wise area graph*, is composed of stacked area charts drawn separately for every age group and assembled on a Lexis-like grid. The fourth is a collection of heatmaps portraying different subsets of the data. The resulting visualization is also known as trellis plot, lattice chart or panel chart. Tufte (1990) refers to it as *small multiples*. This conventional techniques serves as a benchmark to compare our innovations against. Furthermore we propose a slight refinement to the small multiple graph making it more suitable for the display of compositional data.

We assess the quality of the proposed visualizations by discussing various aspects and features of each technique. The different visualizations are compared regarding their data content, their geometrical preservation of the Lexis surface, their ability to communicate precise values as well as general data patterns and their space economy. Some evaluation criteria, like the kind of statistic that is actually displayed in the visualization, can be assessed precisely. Other questions, like the ability of the visualization to show patterns in the data, are ultimately a matter of subjective judgements. We back up our personal judgements of these subjective criteria by references to experiments done in graphical perception and by demonstration of the visualizations using real world data.

The paper consists of 9 sections. In Section 2, we briefly present the Lexis surface, which represents the state-of-the-art of visualization of demographic data grouped by age, period (and cohort). The techniques presented in this paper are extensions of the Lexis surface to the domain of compositional data. Section 3 introduces colour terminology and the concepts of colour composition and colour spaces – matters which are referred to in the subsequent description

---

[2]Ternary refers to a system with three states.

of the visualizations, sections 4–7. Finally we compare the proposed visualization techniques with each other and assess the individual features according to our evaluation criteria.

2. The Lexis Surface

The *Lexis diagram* connects calendar time with age and cohort by the use of a two dimensional diagram with period on the ordinate (*x*) and age on the abscissa (*y*). By interconnecting the three essential time-scales of demography the Lexis diagram serves as an aid in working with population processes and is used to represent events (e. g. birth, death, migration) or occupied states (e. g. single, cohabiting, married, divorced) along the individual life course as well as on the population level. Ages are represented as horizontal parallels, single years as vertical parallels, and each birth cohort is represented by a 45° diagonal. Individuals or populations can be located and tracked on the Lexis diagram as they progresses through time and age (see figure 1). The tool is widely used by demographers and epidemiologists to study population and disease dynamics.[3]
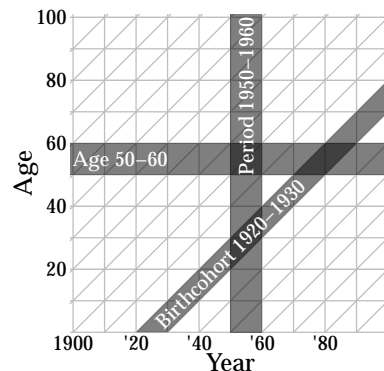

Figure 1: The Lexis diagram.

The term *Lexis surface* has been introduced by Arthur and Vaupel (1984) to describe a collection of demographic rates given by discrete time and age. The term now extends to the visualization of population data (counts, rates) across time and age in the form of contour maps or heatmaps (for examples see e. g. Rau et al. 2008; Scherbov and Vianen 2002). *Contour-maps* indicate regions of equal value for a variable on the time-age plane by a contour line (isoline). Vaupel, Gambill, and Yashin (1987) trace the use of these graphs in Demography back to Kermack, McKendrick, and McKinlay (1934) and Delaporte (1941). *Heatmaps* express the value of a variable for every point on the time-age plane by the use of colour (e. g. light regions might indicate high values, dark regions in turn low values, see figure 2). Gambill and Vaupel facilitated the use of these graphs in demography by developing a plotting software running on personal computers: LEXIS (Gambill and Vaupel 1985) was able to produce *shaded contour maps*, a combination of heatmaps with contour lines, from demographic data. *Thousands of Data at a Glance: Shaded Contour Maps of Demographic Surfaces* showcases the software and the Lexis surface plots across a variety of demographic applications and datasets. Recent refinements to the Lexis surface plot were done by Riffe, who plotted fertility rates structured by period, age *and* cohort resulting in a surface made of triangles instead of rectangles. Riffe also proposed and demonstrated the use of equilateral coordinates to avoid visual distortion along the cohort lines. The programs are written in R and published on Riffes personal webpage (Riffe 2014).

This paper applies the term "Lexis surface" to the visual display of compositional population data given on an period-age-grid.

---

[3]Implementations are available for the statistical software R: The packages Epi (Carstensen et al. 2014) and Biograph (Willekens 2013) both include functions to display data on a time-age-cohort grid.
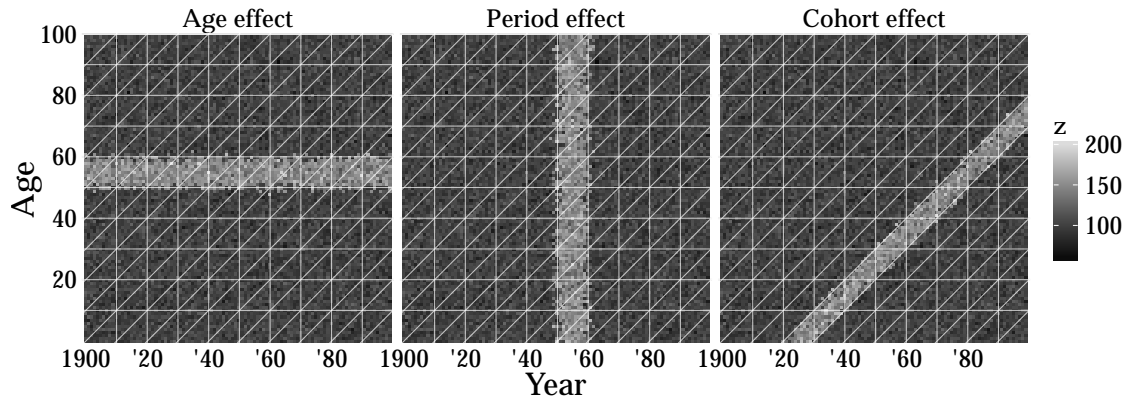
Figure 2: Heatmap representation of Age-, Period-, and Cohort-effects on the Lexis-surface (simulated data).

3. About Colour

How to use colour to visualize compositional data? Two of the visualizations presented in this paper – the ternary-balance-scheme and the qualitative-sequential-scheme – show possible solutions to that question. The ternary-balance-scheme employs the composite nature of mixed colours to display compositional data. The qualitative-sequential scheme maps different dimensions of colour to different data dimensions. In order to understand these techniques some understanding of colour terminology, colour mixing and colour spaces is needed. We will start off with a familiar setting:

Children learn how to produce a wide range of colours by mixing blue, red and yellow paints. The resulting colour depends on the ratio between these *primary colours*: Mix yellow with blue to get green, red with blue to get purple and three parts of yellow with one part of red to produce orange. If you want to change the *lightness* of the colour, add either black (darkens) or white (brightens) paints into the mix. Digital printers use the same principle but use a slightly different set of primary colours, namely cyan, yellow and magenta. A colour computer screen uses yet a different set of primaries: its pixels are composed of red, green and blue subpixels. The relative light-output of these subpixels (*luminance*) determines the colour of the pixel. Furthermore, when mixing red, blue and green light sources by equal parts the resulting colour will be white (*additive colour mixing*) whereas mixing yellow, blue and red paints on a canvas will result in a dark brown (*subtractive colour mixing*).

We can see that there is much more to colour mixing than the few rules taught in basic art classes. The choice of primary colours in the end is arbitrary (Ware 2013:100 f). There is nothing special about blue, red and yellow. Additionally, when working with coloured light sources you will observe different mixtures compared to working with coloured, reflective surfaces.

Colour spaces provide a rigorous way to work with colour. They aim to parametrize and quantify it. Some refer to our experiences of how colour is mixed in everyday situations (RGB for light sources, CYMK for printing), others are more abstract. The common ground is the prediction of a colour dependent on a set of input parameters. In case of the RGB colour space – used to display colours on a television or computer screen – the input parameters are the

4

amounts of red, green and blue light emitted, commonly encoded with values from 0-255. Here the mixed colour is expressed as a composition of primary colours.

It is also possible to build a colour space on parameters relating to human colour perception. When asked to describe a colour it would be a remarkable feast for a person to express it in terms of the proportions of primary colours. Instead we tend to say that a colour is bright or dark (*lightness*), tending to blue, red, green or purple (*hue*) and being very neon or more on the grey side (*chroma*, see Fairchild 2005 chapter 4 for colour terminology).

The *CIE-Lch* (Lightness, chroma, hue) colour space expresses a colour in terms of these three perceptual parameters. Geometrically it can be thought of as a cylinder filled with different colours. The hue of the colours changes around the circumference, the chroma changes along the radius, with neon colours far out and perfect grey at the centre. The lightness changes along the hight of the cylinder with black at the bottom and white at the top. This colour space is not parametrized in terms of primary colours. Therefore colour mixtures are not to be seen from the parameters. However, choosing multiple points in the CIE-Lch colour space, colour mixtures can be constructed by simple geometric operations (see Appendix A).

Apart from an intuitive interpretation of the parameters CIE-Lch has other features making it feasible for use in visualization. Its parameters are not correlated with each other, e. g. changing the lightness will not change the hue or chroma of a colour. Furthermore the function that maps the parameters to the predicted colours is derived from experiments in colour perception. If we specify a range of colours with equal lightness the average observer will perceive the colours as very similar in lightness. On the other hand, the magnitude of perceived differences between two colours will roughly correspond to the magnitude of the quantitative differences in the colour parameters. The usage of such a *uniform* colour space is crucial when translating quantitative data into truthful visual representations (Ware 2013:105, Fairchild 2005:185).

4. Ternary-Balance-Scheme

Cartographers have a long tradition of using colour for data representation and it comes to no surprise that elaborated systematics on the use of colour emerged in this field. A field-agnostic systematic is proposed by Brewer (1994b) (see also Brewer 1994a). Starting with the established qualitative-, binary-, sequential- and divergent colour-scales Brewer also addresses combinations of these scales such as the *Three-variable-balance-scheme*.

The idea is to represent the proportions among three groups with a mixture of three basic colours. The basic colours have different hues, each corresponding to a category of a ternary variable. They are mixed in relations equal to the relations between the three groups in the data. The resulting colour mixture is unique for every possible combination of shares among three groups. The colour-coded ternary diagram can be used as a legend for this type of graph.

The colours in the ternary-balance-scheme follow two basic principles which facilitate intuitive interpretation of the results:

1. The higher the share of a group the more the mixed colour resembles the base colour for that group, and
2. the more equal the group shares are the more the mixed colour tends to grey.

Using colour mixtures to display multiple data dimensions in a single colour has been proposed several times (e. g. Trumbo 1981, Eyton 1984, Ware 1988). The strength in the ternary-balance-scheme lies within the natural interpretation of its colour-mixtures. It is not necessary to memorize the complete legend to make sense of the data. All the reader has to do is to recall the three basic group colours (like in a categorical colour scale) and the interpretation follows the two simple rules given above. With the use of a ternary diagram as legend we are able to colour-code all possible proportions between three groups in a way reminiscent of the representation of colour spaces in two dimensions.

(a) Proportion of people dying from a given cause by time and age (France, total population).

(b) Legend for Ternary-balance-scheme graph of French cause of death data.
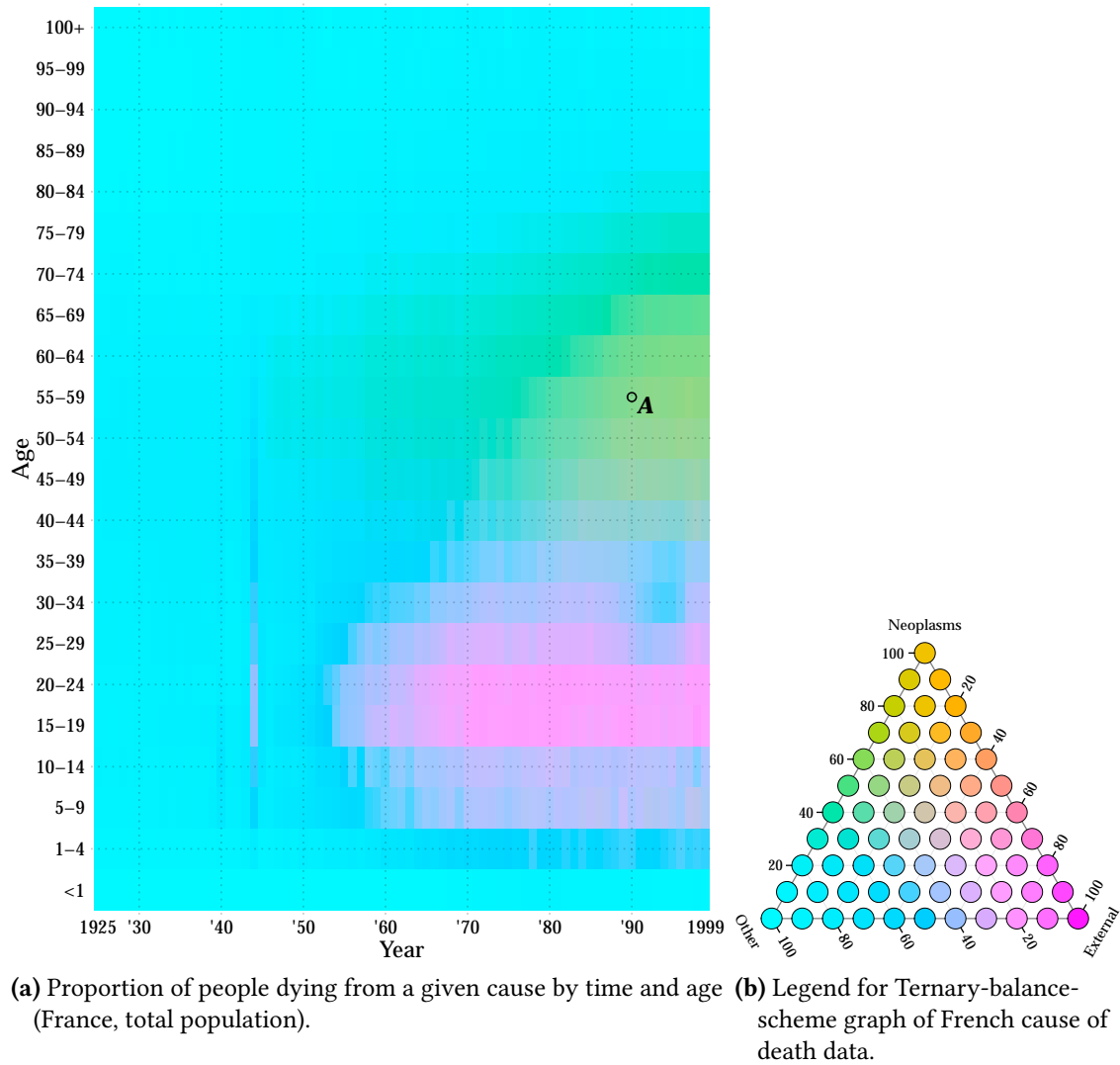
Figure 3: Ternary-balance-scheme: Proportion of people dying from a given cause by time and age.

Figure 3 shows the ternary-balance-scheme applied to age-wise French cause-of-death data across the 20th century. All deaths are divided by cause into the categories *Neoplasms* (ICD-9 codes 140–239), *External* (injury, suicide, accident; ICD-9 codes 800–999 and E–V) and *Other* (all remaining causes of death). Magenta was chosen as a primary colour for external causes of death, orange means death by neoplasm and cyan encodes all remaining deaths.

EXAMPLE: Consider point A in figure 3. The proportion of deaths caused by neoplasm is given in the legend by the position on a horizontal line through the point with the colour of A; likewise the proportion of deaths caused by external causes is indicated by the position on a / line and and the share of all other causes is indicated by the position on a \line. The mixture of magenta, orange and cyan at point A indicates that more than half of the deaths (about 60 %) are caused by neoplasms, 10 % by external causes and the remainder (30 %) by all other causes of death.

The mid-century marks a turning point. Before 1950 external causes of death and cancer were only sporadically found on the death certificates the prominent exception being World War II. Period effects are visible in 1940 (German occupation of France) and – much stronger – in 1944 (Allied landing in Normandy). In both years the war contributed to external causes of deaths, visible in the age range 5–60. After 1950 two major trends in the distribution of death causes emerge: 1. Adolescent deaths rapidly become dominated by external mortality. The "accident hump" (Heligman and Pollard 1980) as a juvenile pattern of mortality comes into existence. 2. Deaths due to cancer gradually become more common in the age range 40–80, dominating the ages 50–70 since the 1980. The onset-age of "other" causes of death as most prominent increases in a linear fashion.

The ternary-balance-scheme allows us to identify major patterns in the change of group-composition over time and age. We are able to identify group-mixtures and group-monopoles. Period-, age- and cohort-effects look much like they would on a 1-dimensional heatmap of all-cause mortality rates on a Lexis surface: Local outliers are visible by a sharp shift in hue – vertical for period effects, horizontal for age effects and on a 45° slope for cohort effects. Slower transitions are visible through smooth colour gradients. The obvious drawback of the ternary-balance-scheme is the limitation to three distinct groups. Also, using colour as a quantifier, small changes in the values are hard to detect and the very nature of the graph makes it unsuitable to use for people with impaired colour-vision (though this limitation could be somewhat relaxed by changing the lightness between the hues).

5. Qualitative-Sequential-Scheme

Another form of multi-variable colour scale is the qualitative-sequential-scheme (Brewer 1994b). The idea is to use a qualitative palette of different hues to signify group-membership and to construct a sequential palette for every group by varying the lightness of the group-colour and thereby encoding group-specific quantities.

See figure 4 for an example using the same dataset as in figure 3 but with two additional categories, namely death due to infectious diseases (ICD-9 codes 001–139) and diseases of the circulatory systems (ICD-9 codes 390–459). Each group is given its own sequential colour scale.

The values given by the colours on the Lexis surface represent the *share of deaths from the most prominent cause of death at year t and age x.*

This approach allows to account for more than three groups in the graphic. The downside is an incomplete picture for every point on the Lexis surface. Only information on the group with the highest share is given. However, in a dataset with a lot of change between the group proportions interesting patterns emerge.

EXAMPLE: Consider the horizontal line at age group 60–64 in figure 4. Before 1945, we see a grey colour: "other" causes of death are dominant. This dominance declines about 1940 (a lighter grey) until, in 1945, circulatory diseases become the main cause of death with a share on all deaths of 20–40 %. Around 1970, the main cause of death shifts from circulatory diseases to neoplasms (the hue shifts from red to yellow). The dominance of neoplasms over other causes of death for 60–64 old males and females increases with time (the yellow hue gets darker and more saturated).

One new insight this visualization yields is the importance of infectious diseases as a cause of death for people aged 15–40 prior to 1950. In these years and ages 20–60 % of the deceased die from infection making this the most likely cause of death in adolescence and middle-age. This pattern abruptly changes after 1950 with external causes of death taking the top spot. Similar to figure 3 we see the accident hump peaking around ages 15–30. Looking at data for five groups we get a more differentiated picture at the death causes in higher ages. Old age mortality is dominated by failures of the circulatory system. The onset of circulatory conditions as the main cause of death moves into higher ages in a linear fashion starting at around age 60 in the 1940s to age 75 in the 1990. For people with an extraordinarily long lifespan the death causes are more diverse with "other" causes of death in the lead.

While confined to visualize shares only for the most prominent group at any given point on the Lexis surface the qualitative-sequential scheme still reproduces a lot of the information gained from figure 3 while adding new insights. In cases where the changes of the group composition in the data are more subtle (no shift in order of group shares) this visualization would loose most of its explanatory power.

6. Age-Wise-Area-Graph

The *stacked-area-chart* can be thought of as the continuous version of the stacked-bar-chart. Coloured areas indicate group shares which change along the x-axis. As the value of each individual group share is indicated by *length* as opposed to colour this technique is better suited to detect slight changes in group composition over time than the previous approaches. Cleveland and McGill (1984) show that people are able to decode visual information more precise into their numerical equivalents when it is represented by length and not by colour.

For the *age-wise*-area-graph we produce the stacked area chart separately for every age-group and assemble all of them on a Lexis-like grid. See figure 5 for the resulting plot. Unlike the ternary-balance-scheme we are able to distinguish more than three groups and unlike the qualitative-sequential-scheme the full information about the group distribution is retained.

EXAMPLE: One feature of the data which was hidden by the former graphical representations is the unusual surge of fatal infectious diseases in ages 25–40 around the mid 1990s. Looking
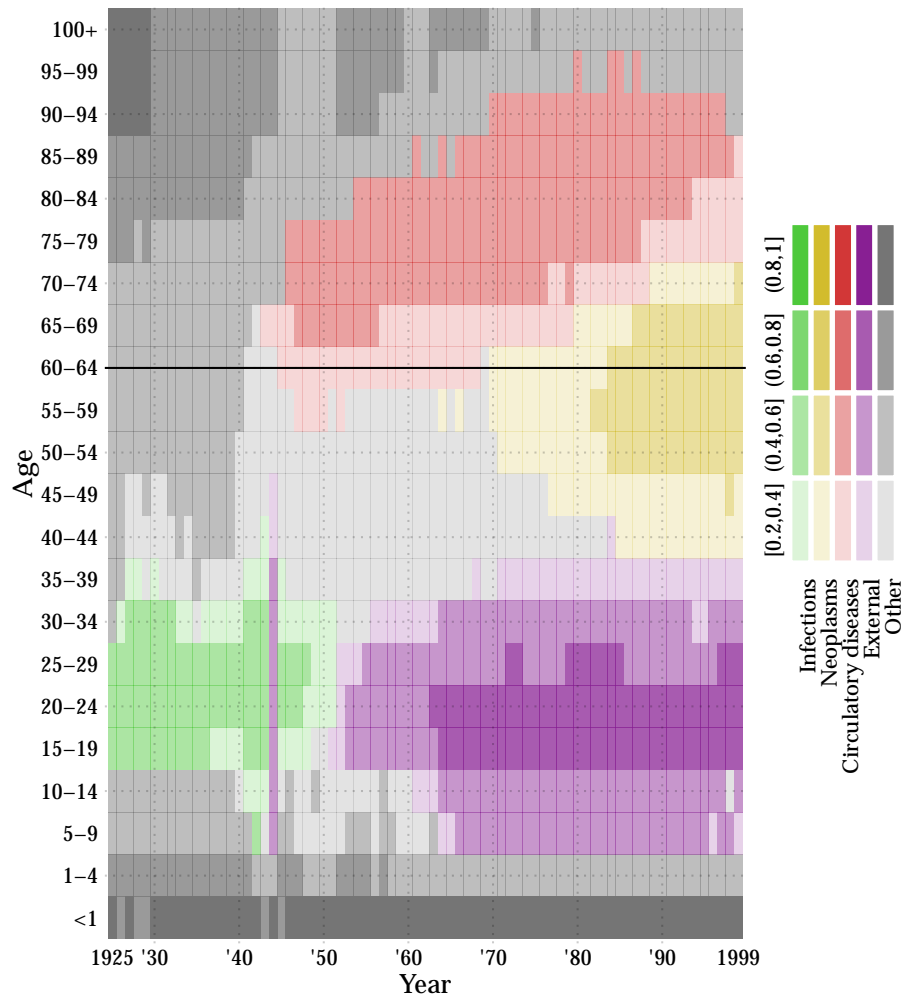
Figure 4: Qualitative-sequential-scheme: Share of people dying from the most prominent cause in a given time and age (France, total population).

at the exact cause of death data we see that HIV related deaths are a cause for this local phenomena. Other new insights from this graph are the relative importance of cancer as a cause for childhood mortality and the high share of "other"-cause-mortality for people aged 90+.

The graph does not exhibit a true Lexis surface. The period-age grid is *non-continuous* as it is composed of multiple stacked-area-charts each having a separate y-axis ranging from 0 (0 %) to 1 (100 %). This breaks the plot area into separate sections making the perception of global patterns harder because the global graphical patterns (the shape of the equal-colour areas across period and age) are interrupted along the age-scale. Also, while area charts (or continuous divided bar charts to stay within the vocabulary of Cleveland) are better than colour encodings for exact table-look up operations they are far from optimal for exact judgements about singular data values (Cleveland 1994).
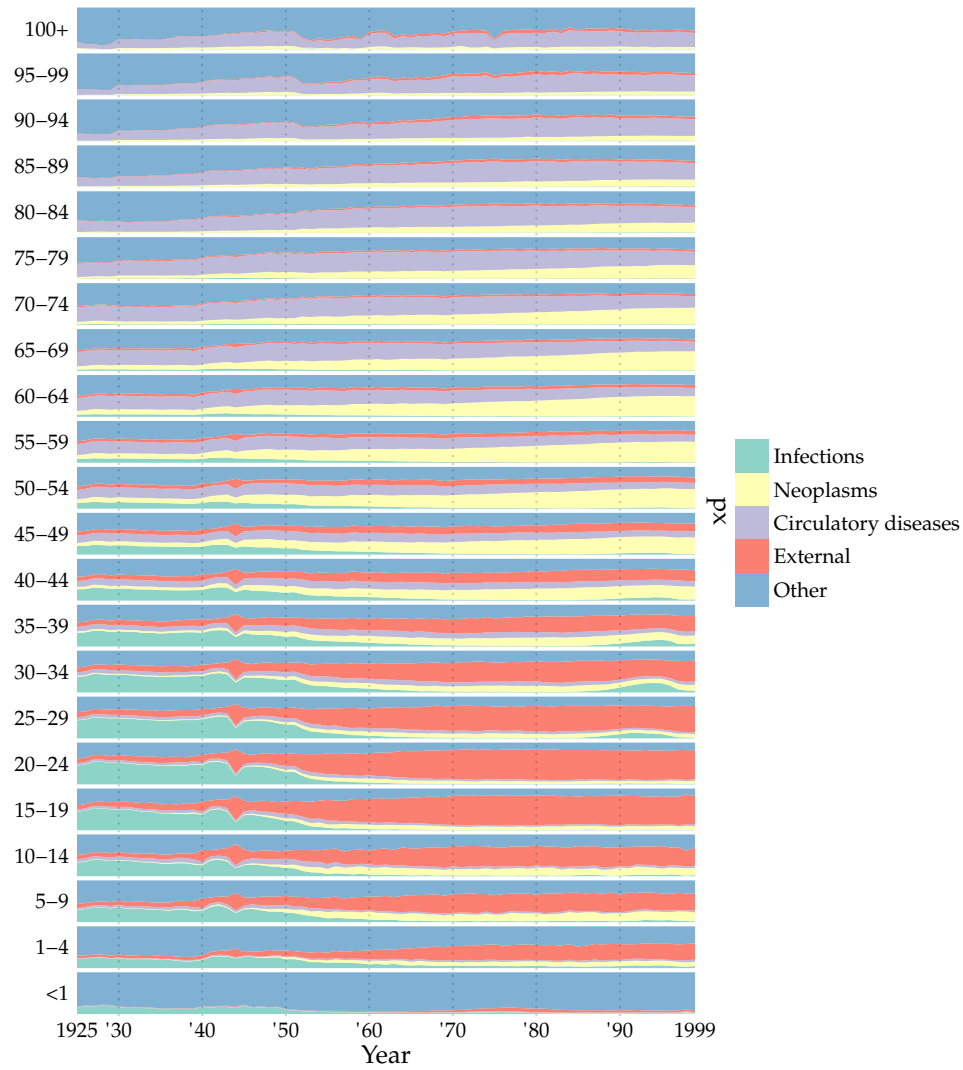
Figure 5: Age-wise-area-graph: Proportion of people dying from a given cause by time and age (France, total population).

The focus of the age-wise-area-graph lies on the detection of small local phenomena and developments within single age groups while still giving an overview of the global patterns for multiple groups on a single period-age-grid.

7. Small-Multiples

Small-multiples[4] are related graphs displayed close to another differing only in the data dimension shown. They can be thought of as "tables of graphs" (Wilkinson 2005:319). Small-multiples are a fairly common technique and have already been discussed by Vaupel, Gambill, and Yashin (1987) as a way of plotting mortality/fertility-surfaces for multiple sub-populations. We discuss this technique in the context of compositional data and propose a small adjustment facilitating cross-panel comparisons.

In figure 6 we see multiple heatmaps of cause-specific shares of deaths in France across period and age. The graphs are augmented by border lines at calendar years and ages for which a given cause of death is dominant.

EXAMPLE: Consider the lower left panel. It shows the proportion of deaths caused by infectious diseases by age and calendar time. Point *A* indicates that in 1930 infectious diseases are the dominant cause of death for persons aged 30-34. Between 40–50 % of all deaths are caused by infection. Infectious diseases stop being the dominant cause of death around 1950. Injuries start to become dominant for persons between 1 and 40 in 1950. This leading position strengthens over time.

Unlike the other visualization techniques described in this paper the small-multiples allow us to display the death-shares for each of the 10 most prominent categories of ICD-9. This power of course comes with a price: The data is not contained in a single graphic. This means that the eye (and the mind) not only have to move between and compare different regions of a graph but make connections between different graphics in order to *see the whole picture*. To make these comparisons easier we mark every period-age combination in which the corresponding cause of death has the highest share among all deaths with a black outline. This way it immediately becomes obvious where specific causes of death dominate over all other causes.

Yet again the new visualization technique shines a different light on our data and allows for additional insights. We can see that ill-defined causes of death are common in the old ages. Prior to World War II an ill-defined cause of death is the norm for people dying at ages 80 or higher. Perinatal conditions followed by congenital anomalies are the main causes of death for infants through most of the 20th century in France. Prior to 1950 deaths of infants and young children are, for the most part, attributed to respiratory diseases. For adults fatal respiratory diseases move into higher and higher ages, reaching a plateau around 1970 and being nearly exclusively an old-age phenomena from there on. A period-age effect can be identified in the realm of digestive diseases. The latter half of the 20th century sees a rise in the share of people aged 35-65 dying from conditions of the digestive system. However, note that the data is binned into 10 % categories simplifying a correct look-up of the colour values but forbidding the exact judgement of small differences.

---

[4]We use the terminology of Tufte (1990). Closely related concepts are *facets* (Wilkinson 2005) or *trellis-plots* (Becker, Cleveland, and Shyu 1996).
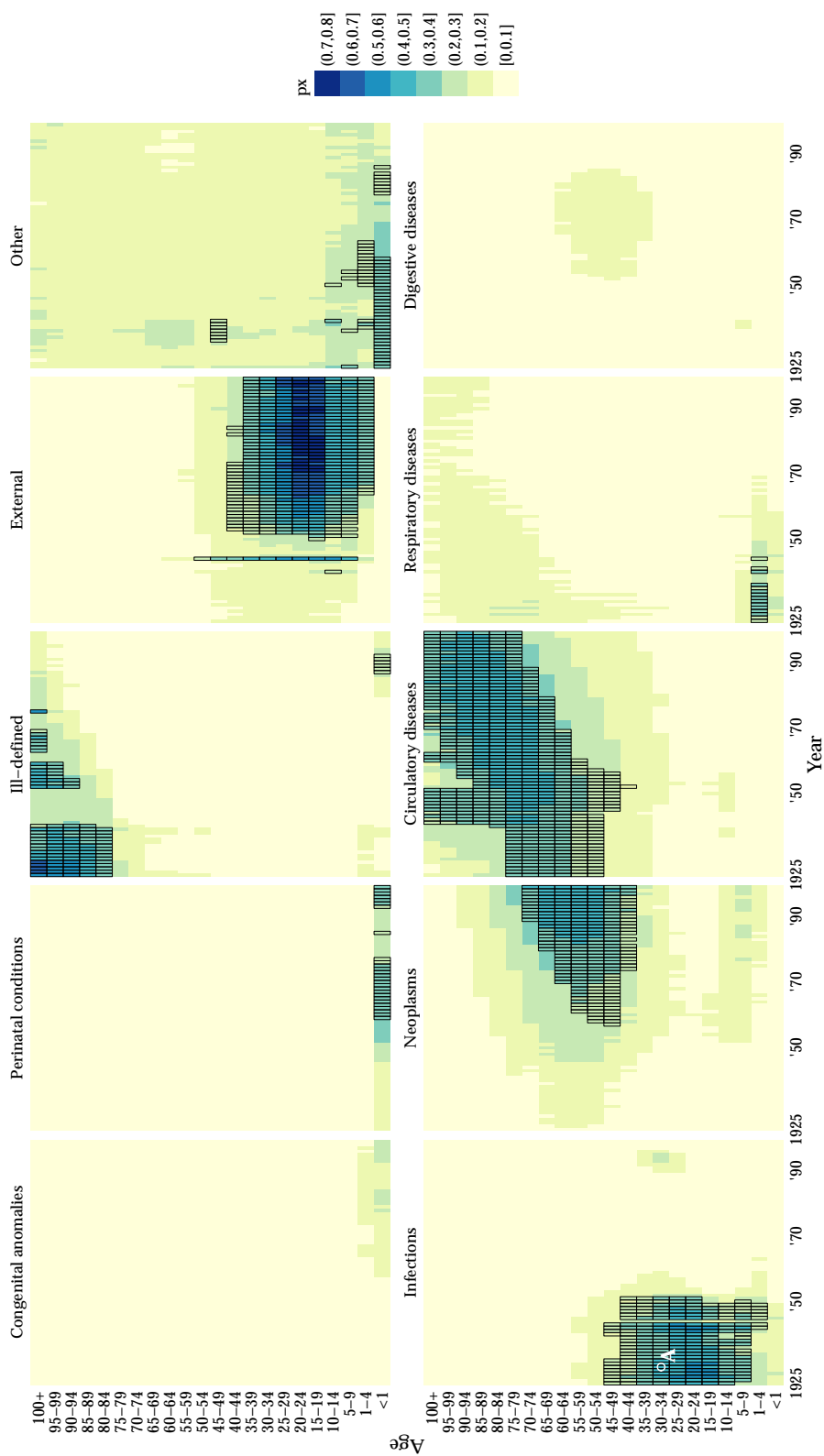
Figure 6: Small-multiples: Proportion of people dying from a given cause by time and age (France, total population).

8. Comparative Evaluation

Completeness is the question whether or not all values of the compositional data (e. g. all individual group shares for each period-age) are displayed in the visualization or if only a subset of the group data (e. g. the share of the largest group for each period-age) is visualized. Visualising the complete dataset is desirable during exploratory data analysis.

If the visualization must contain complete information about every group in the composition the qualitative-sequential-scheme is not an option as only the group with the highest share at a given period-age is considered. The remaining three techniques always are sufficient in showing the complete distribution of group shares.

Continuity requires that the continuous Lexis grid is preserved in the visualization, e. g. the period and age axes constitute a rectangular grid unbroken by whitespace or period-wise/age-wise scales.

The Lexis surface approximates a continuous grid of age by time. The age-wise-area graph breaks this essential feature by fracturing the surface into separate graphs by age. As each age group uses its own y-axis the surface becomes discontinuous along the age dimension. From a perceptional standpoint this hinders the identification of period- and cohort-effects as these would form continuous patterns along the age-(time)-axes but cannot do so in the age-wise-area-graph. The remaining visualization techniques use a continuous Lexis surface.

The Category limit is the maximum number of categories/groups in the compositional data that can be displayed at once. It is precise if limited by technical reasons or approximate if limited by perceptional reasons.

The ternary-balance-scheme – as the name suggests – only allows for the display of proportions between three groups. If a visualization of ratios between a large number of categories is required the small-multiples approach is best suited as a virtually unlimited number of group shares can be displayed at once.

The maximum number of groups allowed for in the qualitative-sequential-scheme and the age-wise-area-graphs is not restricted by technical- but by perceptional limits. The qualitative-sequential scheme uses lightness steps within different hues to show different group shares. Therefore the number of groups which can sensibly be displayed at once is limited by the number of distinct hues one can perceive. Different lightness steps based on these hues should not be confused across hues. Ware (2013):123 lists four unique hues: red, green, blue and yellow. These hues are are generally perceived as pure/unmixed colours which do not resemble each other. Yellow is problematic as a base hue for a lightness sequence as it is already a very light colour (ibid.:127). Orange, still being perceived distinct from red and yellow, is the better choice here. Adding black one can construct five sufficiently different lightness sequences. Some few additional hues, like purple, might also work.

The age-wise area-graphs contain a lot of information in a small vertical space. The more groups one adds the more crowded this vertical space gets and information retrieval will be more difficult.

Pattern perception describes the "visual decoding of physical information" (Cleveland 1994:223 f). Within the context of our paper this translates into the degree at which period-, age-, and cohort-effects in the composition of the data are revealed by the visualisation.

The ternary-balance-scheme, the qualitative-sequential-scheme and the small-multiples are all heatmaps and differ only in the applied colour-scale. As such they are particularly suited to show patterns in the data such as period-, age- and cohort-effects or local outliers (as demonstrated for 1-dimensional data in Vaupel, Gambill, and Yashin 1987). For the small-multiples approach this merit only applies to patter-perception within groups as the information is spread out across multiple panels. The inter-group patterns are not directly visually encoded and have to be inferred from a comparison between multiple panels. The marking of modal values eases this task.

The case for age-wise-area-graphs is more complicated. Due to the discontinuities across the age dimension the global pattern-perception from these graphs is comparatively worse. However, changes over time within single age groups, even small ones, can be seen very well using this approach.[5]

Table look-up "[is the] visual decoding of scale information [...]" (Cleveland 1994:223 f). It takes place when attempting to extract the exact numbers from a displayed data point. Some visual primitives (such as xy-position, height, size etc.) – and therefore some visualizations – make this task easier than others. Being able to easily decode the data values from a visualization is desirable as it allows for numerical judgements (e. g. "*A* is 2.5 times higher than *B*").

None of the discussed techniques allows for a very efficient table look-up operation (the quick and accurate retrieval of the underlying data values). Colour, as used by three out of the four techniques as encoding for the group shares, is generally regarded as the weakest graphical element in terms of table look-up (Cleveland and McGill 1984:536). The judgement of areas and lengths as used in the age-wise-area-graphs fares better, but the matter is complicated by the use of *stacked-relative*-area-graphs. Here the value can only be directly read from the y-axis for the bottom group/area. For all other areas differences between two points on the y-axis equal the data value.

The Footprint is often of concern when publishing results. Some visualizations gain their descriptive power by occupying a large area.

The footprint of all techniques but the small-multiples is equal to a conventional heatmap across the same period-age-range. The small-multiples technique, depending on the number of groups, might be much larger up to the point of filling the whole page.

---

[5]Given that no steep slopes occur or else there will be problems correctly judging the difference between the upper and lower boundary of an area – the value (Cleveland 1994:227 ff).

	Ternary-balance-scheme	Qualitative-sequential-scheme	Age-wise-area-graph	Small-multiples
Completeness	yes	no	yes	yes
Continuity	yes	yes	no	yes
Category limit	3	≈ 5–6	≈ 8	unlimited
Pattern perception	good	good	limited	good
Table look-up	limited	limited	limited	limited
Footprint	small	small	small	large

Table 1: Evaluation of different visualization techniques for compositional data on the Lexis surface.

9. Conclusion

We demonstrated four different visualization techniques for showing cause of death distributions across period and age, extending the conventional period-age heatmap of 1-dimensional continuous data (e.g. surfaces of mortality rates) to multidimensional compositional data. We applied multivariate colour-scales originating from cartography to the Lexis surface (ternary-balance-scheme, qualitative-sequential-scheme), introduced age-wise area graphs as a means of visualizing compositional change over time within ages and improved upon the well known small multiple technique in the context of compositional data. Each of these techniques serves the cause of making sense of compositional data across time and age, while at the same time these techniques complement each other by having different strong points.

The ternary-balance-scheme is the best candidate for visualizing the *shares of three groups.* Using principles of colour composition it produces smooth rainbow-like surfaces immediately indicating time-, age-, and cohort patterns in the data. The qualitative-sequential-scheme *extends beyond three groups* and shows the shares of the most dominant groups on a surface. The age-wise-area-graphs use the power of the line to point out *slight, age-specific changes* in group composition over time. Plotting a separate heatmap for each group is a well-tried technique and still the most practical way to plot compositions for a *large number of groups.* Adding a contour line to each panel, indicating the regions of dominance of one group over the others, reduces the need for cross-panel comparisons and therefore adapts this proven technique further to the display compositional data.

What share of a population features attribute i at period t and age x? The proposed visualizations are tools helping to answer this question. We demonstrated the techniques using data on causes of death. Other possible applications include time-age surfaces of population shares by labour-market status (unemployed and seeking job, unemployed and not seeking job, employed), distributions of labour force over industries (agricultural, industry, service) or partnership status (single, in relationship and living alone, in relationship and cohabiting, married). The visualizations can also help interpreting the output from estimated models. The Heligman and Pollard model of overall life course mortality (Heligman and Pollard 1980) for example is written as the sum of age-specific components for childhood-, accident- and senescent mortality. The relative magnitudes of these components can be clearly visualized using the ternary

balance scheme, producing graphics of quasi cause-specific-mortality from all cause mortality data.

Good visualizations can be thought of as a visual model of the data at hand, able to identify relationships between variables. They go hand in hand with mathematical models of the data as both can be checked against the other. We hope to have contributed useful techniques for revealing information about compositions across time and age.

References

Aitchison, John (1986). *The Statistical Analysis of Compositional Data.* Ed. by John Aitchison. Monographs on statistics and applied probability. New York: Chapman and Hall.

– (2002). "Biplots of compositional data". In: *Applied Statistics* 51.4, pp. 375–392.

Arthur, W. Brian and James W. Vaupel (1984). "Some general relationships in population dynamics". In: *Population Index* 50.2, pp. 214–226.

Becker, R. A., William S. Cleveland, and M-J Shyu (1996). "The design and control of Trellis display". In: *Journal of Computational and Graphical Statistics* 5, pp. 123–155.

Brewer, Cynthia A. (1994a). "Color Use Guidelines for Mapping and Visualization". In: ed. by Alan M. MacEachren and D. R. Fraser Taylor. Vol. 2: Visualization in Modern Cartography. Modern Cartography Series. Academic Press. Chap. 7, pp. 123–147.

– (1994b). "Guidelines for use of the perceptual dimensions of color for mapping and visualization". In: *SPIE* 2171, pp. 54–63.

– (1999). "Color Use Guidelines for Data Representation". In: *Proceedings of the Section on Statistical Graphics.* American Statistical Association. Alexandria, VA, pp. 55–60.

Carstensen, Bendix et al. (2014). *Epi: A package for statistical analysis in epidemiology.* Online 2014–11–26. The R Project for Statistical Computing. URL: http://cran.r-project.org/web/packages/Epi/index.html.

Cleveland, William S. (1994). "The Elements of Graphing Data". In: ed. by William S. Cleveland. Hobart Press. Chap. 4, pp. 221–270.

Cleveland, William S. and Robert McGill (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387, pp. 531–554.

d'Angeville, Adolphe, ed. (1836). *Essai sur la Statistique de la Population francaise.* Bourg-en-Bresse: F. Doufour.

Delaporte, P. (1941). *Evolution de la mortalité en Europe depuis l'origine des statistiques de l'Etat civil (Tables de mortalité de generations).* Imprimerie Nationale.

Eyton, J. Ronald (1984). "Complementary-Color, Two-Variable Maps". In: *Annals of the Assiciation of American Geographers* 74.3, pp. 477–490.

Fairchild, Mark D. (2005). *Color Appearance Models.* Ed. by Michael A. Kriss. 2nd ed. Wiley – IS&T Series in Imaging Science and Technology. Chichester: Wiley.

Gabriel, K. R. (1971). "The biplot graphic display of matrices with application to principal component analysis". In: *Biometrika* 58.3, pp. 453–467.

Gambill, Bradley A. and James W. Vaupel (1985). *The LEXIS program for creating shaded contour maps of demographic surfaces.* Working Paper WP-85-94. Laxenburg, Austria: International Institute for Applied Systems Analysis (IIASA).

Heligman, L. and J.H. Pollard (1980). "The Age Pattern of Mortality". In: *Journal of the Institute of Actuaries* 107.1, pp. 49–80.

Kermack, W., A. McKendrick, and P. McKinlay (1934). "Death-rates in Great Britain and Sweden: some general regularities and their significance". In: *The Lancet* 31, pp. 698–703.

Lexis, Wilhelm Hector Richard Albrecht (1875). *Einleitung in die Theorie der Bevölkerungsstatistik.* Ed. by Wilhelm Hector Richard Albrecht Lexis. Straßburg: K. J. Trübner.

Rau, Roland et al. (2008). "Continued Reductions in Mortality at Advanced Ages". In: *Population and Development Review* 34.4, pp. 747–768.

Riffe, Tim (2014). *Tim Riffe Personal*. Online 2014–11–26. URL: https://sites.google.com/site/timriffepersonal/.

Scherbov, Sergei and Harrie van Vianen (2002). "Period Fertility in Russia since 1930: an application of the Coale-Trussell fertility model". In: *Demographic Research* 6, pp. 455–470.

Trumbo, Bruce E. (1981). "A Theory for Coloring Bivariate Statistical Maps". In: *The American Statistician* 36.4, pp. 220–226.

Tufte, Edward R. (1990). *Envisioning Information*. Ed. by Edward R. Tufte. 10th ed. Cheshire, Conneticut: Graphics Press.

Vallin, Jacques and France Meslé (2014). *Database on causes of death in France from 1925 to 1999*. Online 2014–08–19. Institut national d'études démographiques. URL: http://www.ined.fr/en/resources_documentation/detailed_data/death_causes_since_1925/.

Vandeschrick, Christophe (2001). "The Lexis diagram, a misnomer". In: *Demographic Research* 4, pp. 97–124.

Vaupel, James W., Bradley A. Gambill, and Anatoli I. Yashin (1987). *Thousands of Data at a Glance: Shaded Contour Maps of Demographic Surfaces*. Research Report RR-87-16. Laxenburg, Austria: International Institute for Applied Systems Analysis (IIASA).

Walker, Francis A., ed. (1874). *Statistical Atlas of the United States, Based on the Results of Ninth Census, 1870, with Contributions from Many Eminent Men of Science and Several Departments of the Federal Government*. New York: Julius Bien.

Ware, Colin (1988). "Using Color Dimensions to Display Data Dimensions". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 30.2, pp. 127–142.

– (2013). *Information Visualization. Perception for Design*. Ed. by Meg Dunkerley and Heather Scherer. 3rd ed. Waltham, MA: Elsevier.

Wilkinson, Leland (2005). *The Grammar of Graphics*. Ed. by Leland Wilkinson. 2nd ed. Statistics and Computing. New York: Springer.

Willekens, Frans (2013). *Biograph: Explore life histories*. Online 2014-11-26. The R Project for Statistical Computing. URL: http://cran.r-project.org/web/packages/Biograph/index.html.

A. Construction of the ternary-balance-scheme

There are multiple ways to construct the *Ternary-balance-scheme.* To derive the scheme from a perceptual colour-space ensures that differences in the numeric data are mapped to perceptually equivalent differences in the colour representation. The perceived difference between any two hues should correspond to the difference between the underlying data values. Additionally a constant lightness of the colour-mixes is desirable as colour-lightness usually is assigned to the magnitude of a continuous variable (Brewer 1994a), but we are dealing with compositional data which always sums up to the magnitude 1. The *CIE-Lch* colour-space suits these needs as it is perceptually balanced and allows for the truly[6] independent specification of hue (h), chroma (c) and lightness (L) of a colour. The colour space can be thought of as a cylinder with different hues along the circumference, decreasing chroma along the radius towards the centre and increasing lightness along the vertical axis. Figure 7a shows a "slice" of this cylinder with a fixed lightness level. We will use this slice to construct the ternary-balance scheme for our graphic: In a first step the three groups in the data are assigned to *equidistant* hues along the circumference of the circle. These are the base colours for each category. Next vectors originating from the centre and pointing towards the base colours are constructed. The length of each vector is given by the share of the corresponding group on the total. This share is then mapped to the available chroma range (in our example $[0, 140]$, other upper limits are possible) with $c = 0$ corresponding to a share of 0% and the upper limit of c corresponding to 100%. Finally the group specific vectors are added and the resulting vector returns the hue and chroma of the colour-mixture used to represent the distribution of the three groups. The possible colour-mixes form a triangular subset of the original CIE-Lch slice and therefore can be conveniently labelled with a ternary scale which allows for numerical interpretation of each mixed-colour (see figure 7).

(a) CIE-Lch colour-space slice ($L = 80$).

(b) Colour-mixes for different compositions of three groups derived from CIE-Lch.

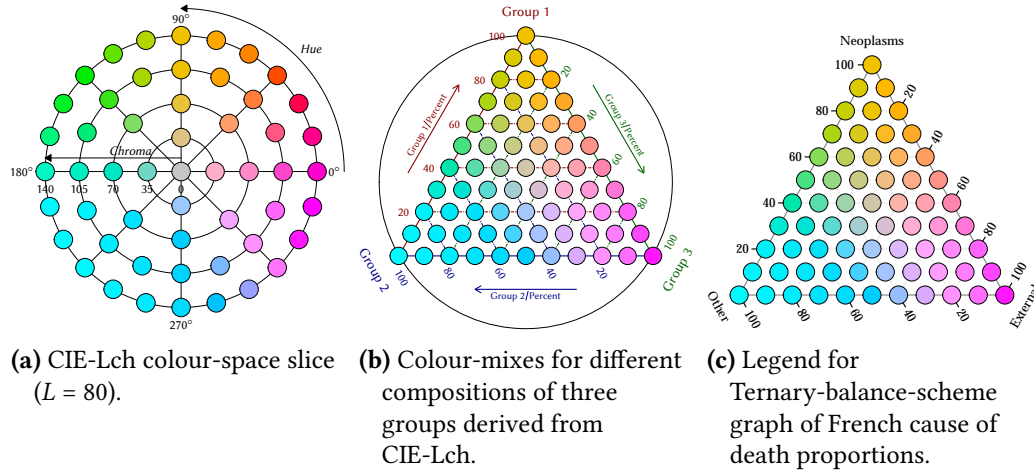**(c)** Legend for Ternary-balance-scheme graph of French cause of death proportions.

Figure 7: Construction of the Ternary-balance-scheme.

---

[6]Unlike the widely used HSV/HLS colour-spaces which confound the perceptual dimensions. Using these colourspaces an increase in saturation also changes the perceived lightning of the colour (Brewer 1999).

B. R code

──────────────────────────── **Script 1** Initialization ────────────────────────────

```
   # Packages -------------------------------------------------

   # visualization
   library(gridBase)      # integrating grid and base graphics
 5 library(gridExtra)     # extended grid functions
   library(ggplot2)       # 2d plotting framework
   library(gtable)        # work with ggplot objects
   #library(ggtern)        # ternary diagrams
   library(scales)        # different scales, RGB-alpha functions
10 library(colorspace)    # handle different colour-spaces
   library(rcpal)         # colour palettes
   library(ggtheme)       # flexible minimal ggplot theme (github)

   # data transformation
15 library(tidyr)         # tidy data, convert between long and wide
   library(dplyr)         # data verbs, operations on subsets of data

   # misc
   library(devtools)      # developer tools, github interface
20 library(colorspace)    # handle different colour-spaces

   # System ---------------------------------------------------

   # System info
25 sessionInfo()

   # Defaults -------------------------------------------------

   # 5 level qualitative colour scheme
30 cpal_qual_5 <- c(rcpal$quacla[4], # green
                    rcpal$quacla[3], # yellow
                    rcpal$quacla[1], # red
                    rcpal$quacla[5], # purple
                    "#757575")       # grey
35

   # font family
   font_family <- "Palatino"
   font_size <- 11
```

———— Script 2 Utility Functions ————

```r
# Util -----------------------------------------------------

#' Export Graphical Object as PDF
#'
#' @param .x Graphical object
#' @param .path Filesystem destination to save object.
#'    [string, length == 1]
#' @param .width Figure width in cm. [numeric, length == 1]
#' @param .height Figure height in cm. [numeric, length == 1]
#'
#' @return PDF output to disk.
ExportPDF <- function (.x, .path, .width, .height) {
  pdf(.path, width = 0.4*.width, height = 0.4*.height,
      useDingbats = FALSE) # avoid problems with missing fonts
  grid.newpage()
  vp <- viewport(x = 0.5, y = 0.5,
                 width = unit(.width, "cm"),
                 height = unit(.height, "cm"))
  pushViewport(vp)

  print(.x, vp = vp)
  dev.off()
}


#' Convert RGB + Alpha Specification to RGB
#'
#' @details Assuming alpha blending with white background.
#'
#' @param .rgb RGB value in hex-code. [character, length >= 1]
#' @param .alpha Alpha value within [0,1]. [numeric, length >= 1]
#'
#' @return RGB-hex value of RGB + alpha after blending with white
#'    background.
AlphaRGBToRGB <- function (.rgb, .alpha) {
  # convert colour hex to (s)rgb
  rgb <- col2rgb(.rgb)
  srgb <- rgb / 255
  # mix colours
  result <-
    mixcolor(.alpha, # mix according to these alpha
             sRGB(1, 1, 1), # mix with white background
             sRGB(srgb[1], srgb[2], srgb[3]))

  return(hex(result))
}
```

iii

```
   # Polar coordinates -------------------------------------------

   #' Sin Using Degrees
   #'
5  #' @param x x in degrees. [numeric, length == 1]
   #'
   #' @return y = sin(x).
   SinD <- function (.x) {
     result <- sin(.x*pi/180)
10   result[.x == 180 | .x == 360] <- 0

     return(result)
   }

15 #' Cos Using Degrees
   #'
   #' @param x x in degrees. [numeric, length == 1]
   #'
   #' @return y = cos(x).
20 CosD <- function (.x) {
     result <- cos(.x*pi/180)
     result[.x == 90 | .x == 270] <- 0

     return(result)
25 }

   #' Convert Polar Coordinates to Cartesian Coordiantes
   #'
   #' @param .r Polar radius. [numeric, length >= 1]
30 #' @param .theta Polar angle in degrees. [numeric, length >= 1]
   #'
   #' @return Cartesian coordinates xy.
   Pol2Cart <- function (.r, .theta) {
     x <- .r*CosD(.theta)
35   y <- .r*SinD(.theta)

     return(cbind(x = x, y = y))
   }

40 #' Convert Cartesian Coordinates to Polar Coordinates
   #'
   #' @param x Cartesian x position. [numeric, length >= 1]
   #' @param y Cartesian y position. [numeric, length >= 1]
   #'
45 #' @return Polar coordinates r (radius) and theta (angle).
   Cart2Pol <- function (.x, .y) {
```

```
        r <- sqrt(.x^2 + .y^2)
        theta <- atan2(.y, .x)
        theta <- theta*180/pi # convert to degrees
50
        return(cbind(r = r, theta = theta))
      }


      #' Return Angles of Equidistant Points on a Circle Perimeter
55    #'
      #' @param .n Number of points. [numeric, length == 1]
      #' @param .start Perimeter position of first point in degrees.
      #'    [numeric, length == 1]
      #'
60    #' @return Position of n equidistant points on circle perimeter in
      #'    angular degrees.
      CircEquiDist <- function (.n, .start) {
        dist <- 360 / .n
        seqn <- cumsum(c(0, rep(dist, .n - 1))) + .start
65      seqn[seqn >= 360] <- seqn[seqn >= 360] - 360

        return(seqn)
      }


70    #' Add Vectors in Polar Coordinates
      #'
      #' @details The origin of the vectors is assumed to be (0, 0). The
      #'    vectors are added in cartesian coordinates with the results being
      #'    converted back to polar.
75    #'
      #' @param .h Hue as theta, polar direction in degrees.
      #'    [numeric, length >= 1]
      #' @param .c Chroma as r, polar radius. [numeric, length >= 1]
      #'
80    #' @return Sum of vectors in polar coordinates (r, theta).
      AddPolVec <- function (.h, .c) {
        cart <- Pol2Cart(.r = .c, .theta = .h)
        cart_sum <- apply(cart, 2, sum)
        pol_sum <- Cart2Pol(.x = cart_sum[1], .y = cart_sum[2])
85
        # handle floating point errors by rounding to next whole number
        pol_sum <- round(pol_sum)

        # make sure the output doesn't use negative degrees
90      pol_sum[pol_sum[,2] < 0, 2] <- pol_sum[pol_sum[,2] < 0, 2] + 360

        return(pol_sum)
      }
```

```
   #' Translate Lexis Group Compositions Into Ternary Balance Scheme
   #'
   #' @param .x x-axis values, usually calendar time. [numeric, length >=
   #'   1]
 5 #' @param .y y-axis values, usually age. [numeric, length >= 1]
   #' @param .group Group affiliation. [factor, length == .x*.y]
   #' @param .p Group share on total. [numeric, length == .group]
   #' @param .colour Turn the colour wheel, use angular degrees as input.
   #'   [numeric, length == 1]
10 #' @param .L Lightness level. [numeric, length == 1]
   #' @param .C Maximum chroma (saturation) level, used in case of
   #'   complete dominance by a single group. [numeric, length == 1]
   #'
   #' @return Ternary balance scheme colours for each xy-position.
15 MixTernBalance <- function (.x, .y, .group, .p, .colour = 90, .L = 80, .C =↙
     ↳ 140) {

     # assemble args into data frame
     simp <- data_frame(Year = .x, Age = .y, Group = .group, p = .p)

20   # generate base colours for each group
     simp %>%
       # set base hue for each group
       mutate(H = HueToGroup(.group = Group, .start = .colour)) %>%
       # set chroma according to the group shares range: [0, C]
25     mutate(C = p * .C) %>%
       # get mixed colors for each year-age
       # mix group colours using vector addition in polarLAB coordinates
       group_by(Year, Age) %>%
       do(data.frame(AddPolVec(.$H, .$C))) %>% ungroup %>%
30     # convert to hex-rgb
       mutate(ID = seq_along(Year),
              RGB = hex(polarLAB(L = .L, C = r, H = theta), fixup = TRUE)) -> ↙
                ↳ result

     return(result)
35 }

   #' Construct Data for Ternary Balance Scheme Legend from 2-Simplex
   #'
   #' @param .simp 2-Simplex as returned by Simp2()
40 #' @param .colour Turn the colour wheel, use angular degrees as input.
   #'   [numeric, length == 1]
   #' @param .L Lightness level. [numeric, length == 1]
   #' @param .C Maximum chroma (saturation) level, used in case of
   #'   complete dominance by a single group. [numeric, length == 1]
```

```r
45   #'
     #' @return Ternary balance scheme colours for each position in ternary
     #'   diagram by steps of 0.1.
     LgndTernBalance <- function(.simp, .colour = 90, .L = 80, .C = 140) {
       .simp %>%
50       # long format
         gather(Group, p, -ID) %>%
         # chroma
         mutate(C = p * .C) %>%
         # hue
55       mutate(H = HueToGroup(.group = Group, .start = .colour)) %>%
         # mixed colours
         group_by(ID) %>%
         do(data.frame(AddPolVec(.$H, .$C))) %>% ungroup %>%
         mutate(ID = seq_along(ID),
60               RGB = hex(polarLAB(L = .L, C = r, H = theta), fixup = TRUE)) %>%
         # join with original simplex
         left_join(.simp) -> result

       return(result)
65   }


     #' Generate a 2-Simplex for Use in Ternary Balance Legend
     Simp2 <- function () {

70     # all combinations of 3 numbers from set:
       # {0, 0.1, 0.2, ..., 0.9, 1}
       expand.grid(x = seq(0, 1, 0.1),
                   y = seq(0, 1, 0.1),
                   z = seq(0, 1, 0.1)) %>%
75       # sum of individual combinations
         mutate(Sum = rowSums(.)) %>%
         # all combinations which sum is 1 (rounding to avoid FP error)
         filter(round(Sum, 3) == 1) %>% select(-Sum) %>%
         mutate(ID = seq_along(x))-> simp_2
80
       return(simp_2)
     }


     #' Assign an Angular Hue Value to Each Group
85   #'
     #' @param .group Group affiliation. [factor, length >= 1]
     #' @param Perimeter position of first point in degrees.
     #'   [numeric, length == 1]
     HueToGroup <- function (.group, .start = 0) {
90     group_lvl <- levels(.group)
       n <- length(group_lvl)
       base_h <- CircEquiDist(.n = n, .start = .start)
```

```
      h <- rep(NA, length(.group))
95    for (i in group_lvl) {
        h[.group %in% i] <- base_h[group_lvl %in% i]
      }

      return(h)
100   }
```

```r
    # ICD-9 codebook -------------------------------------------

    read.csv("./data/cod/cod_names.csv") %>%
      as_data_frame -> cbook_cod
 5
    # INED COD -------------------------------------------------

    # age levels in correct order
    lev_age <- c("<1","1-4","5-9","10-14","15-19","20-24","25-29",
10              "30-34","35-39","40-44","45-49","50-54","55-59",
                "60-64","65-69","70-74","75-79","80-84","85-89",
                "90-94","95-99","100+")

    # cause of death data
15  read.csv("./data/cod/ined-cod-fra-1925-1999-counts.csv") %>%
      as_data_frame %>%
      # read in data on death counts and do basic tidying
      filter(Age != "Total") %>% droplevels %>%
      mutate(Age = factor(Age, levels = lev_age, ordered = TRUE)) %>%
20    mutate(COD = factor(COD, labels = cbook_cod$Short)) %>%
      # convert counts to cause specific shares on total deaths
      group_by(Year, Age, Sex) %>%
      select(Year, Age, Sex, COD, Dx) %>%
      mutate(px = Dx / Dx[COD == "Total"]) %>%
25    filter(COD != "Total") %>% droplevels-> counts

    # aggregate missing shares in "Other" value
    counts  %>%
      group_by(Year, Age, Sex) %>%
30    summarise(COD = "Other", px = 1 - sum(px)) -> temp
    # merge back into original data
    rbind(select(counts, -Dx), temp) %>%
      as_data_frame %>%
      arrange(Year, Age, Sex) -> counts; rm(temp)
35
    # Top 10 CODs ----------------------------------------------

    # share by COD top 10 ICD codes
    lab_cod_10 <- as.character(cbook_cod$Short[c(2, 3, 9, 10, 11, 16, 17, 18, ↙
      ↳ 19)])
40  counts  %>%
      group_by(Year, Age, Sex) %>%
      filter(!(COD %in% lab_cod_10)) %>%
      summarise(COD = "Other", px = sum(px)) -> temp
    # merge back into filtered data
45  counts %>%
```

```
      filter(COD %in% lab_cod_10) -> counts_10
   rbind(counts_10, temp) %>%
      as_data_frame %>%
      droplevels %>%
50    arrange(Year, Age, Sex) -> counts_10; rm(temp)


   # Top 5 CODs -------------------------------------------------

   # percentage by COD top 5 ICD codes
55 lab_cod_5 <- as.character(cbook_cod$Short[c(2, 3, 9, 19)])
   counts  %>%
      group_by(Year, Age, Sex) %>%
      filter(!(COD %in% lab_cod_5)) %>%
      summarise(COD = "Other", px = sum(px)) -> temp
60 # merge back into filtered data
   counts %>%
      filter(COD %in% lab_cod_5) -> counts_5
   rbind(counts_5, temp) %>%
      as_data_frame %>%
65    droplevels %>%
      arrange(Year, Age, Sex) -> counts_5; rm(temp)


   # Top 3 CODs -------------------------------------------------


70 # percentage by COD top 3 ICD codes
   lab_cod_3 <- as.character(cbook_cod$Short[c(3, 19)])
   counts  %>%
      group_by(Year, Age, Sex) %>%
      filter(!(COD %in% lab_cod_3)) %>%
75    summarise(COD = "Other", px = sum(px)) -> temp
   # merge back into filtered data
   counts %>%
      filter(COD %in% lab_cod_3) -> counts_3
   rbind(counts_3, temp) %>%
80    as_data_frame %>%
      droplevels %>%
      arrange(Year, Age, Sex) -> counts_3; rm(temp)
```

```
     # Colour mixing -----------------------------------------------

     # share of deaths [0,1] by cause over Year, Age and Sex
     counts_3 %>%
5      # reorder factors to match specific colours with specific CODs
       mutate(COD = factor(COD, c("Neoplasms", "Other", "External"))) %>%
       # subset to Female & Male
       filter(Sex == "total") %>%
       # do the ternary balance scheme colour mixing...
10     do(MixTernBalance(.x = .$Year, .y = .$Age, .group = .$COD, .p = .$px)) ->↵
         ↳  counts_3_mix

     # data for legend
     Simp2() %>%
       do(LgndTernBalance(.)) -> tern_lgnd
15
     # Plot ternary balance scheme ----------------------------------

     plot_tern_balance <-
       ggplot(counts_3_mix) +
20     # main
       geom_tile(aes(x = Year, y = Age, fill = factor(ID))) +
       # scale
       scale_fill_manual(values = counts_3_mix$RGB) +
       scale_x_continuous(breaks = c(1925, seq(1930, 1990, 10), 1999),
25                        labels = c(1925, "'30", "'40", "'50",
                                          "'60", "'70", "'80", "'90", 1999),
                          expand = c(0.005, 0)) +
       # coord
       coord_fixed(ratio = 5) + # age-year := period year
30     # annotation
       annotate("point", x = 1990, y = "55-59", shape = 21) +
       annotate("text", x = 1990, y = "55-59", label = "italic(A)",
               family = font_family, size = 4,
               hjust = -0.3, vjust = 1, parse = TRUE) +
35     geom_vline(xintercept = seq(1930, 1999, 10),
               colour = "black", size = 0.4, alpha = 0.2, lty = 3) +
       geom_hline(yintercept = seq(2, 22, 2),
               colour = "black", size = 0.4, alpha = 0.2, lty = 3) +
       # theme
40     ggtheme_min(base_size = font_size, base_family = font_family, grid = "n")↵
         ↳  +
       theme(legend.position = "none")

     ExportPDF(plot_tern_balance, "./fig/plot-tern_balance_no_lgnd.pdf", 13, 17)
```

```
45  # Plot ternary balance scheme legend ---------------------------

    # plot legend
    plot_tern_balance_lgnd <-
      ggplot(tern_lgnd, aes(x = x, y = y, z = z)) +
50    # main
      geom_point(aes(fill = as.factor(ID)), size = 8, pch = 21) +
      # coord
      coord_tern(L = "y", T = "x", R = "z") +
      # scale
55    scale_fill_manual(values = tern_lgnd$RGB, guide = FALSE) +
      scale_L_continuous("Other") +
      scale_T_continuous("Neoplasms") +
      scale_R_continuous("External") +
      # theme
60    theme_bw() +
      theme_hidearrows() +
      theme(panel.grid.tern.minor       = element_blank(),
            axis.tern.text              = element_text(size = 12),
            axis.tern.ticklength.major  = unit(0.5, "cm"),
65          axis.tern.title.T           = element_text(vjust = -1),
            axis.tern.title.L           = element_text(hjust = 0.5, vjust = ↵
                ↳ 2.5, angle = -60),
            axis.tern.title.R           = element_text(hjust = 0.5, vjust = ↵
                ↳ 2.5, angle = 60),
            text                        = element_text(family = font_family, ↵
                ↳ size = font_size),
            plot.background             = element_rect(fill = "transparent", ↵
                ↳ colour = NA))
70
    ExportPDF(plot_tern_balance_lgnd, "./fig/plot-tern_balance_lgnd.pdf", 13, ↵
        ↳ 13)
```

```
   # Plot qualitative sequential scheme ----------------------------

   # subset to modal cause of death for each year, sex, age
   counts_5 %>%
5    group_by(Year, Sex, Age) %>%
     filter(px == max(px), Sex == "total") %>%
     ungroup %>% droplevels -> counts_5_mode

   # minimum alpha level (used for lowest share category)
10 alpha_min <- 0.2

   # plot qualitative sequential scheme
   plot_qual_seq <-
     ggplot(counts_5_mode) +
15   # avoid background shine through transparency
     geom_tile(aes(x = Year, y = Age), fill = "white") +
     # main
     geom_tile(aes(x = Year, y = Age, fill = COD,
                   alpha = cut_interval(px, length = 0.2)),
20           colour = "transparent") +
     # annotate
     geom_vline(xintercept = seq(1930, 1999, 10),
                colour = "black", size = 0.4, alpha = 0.2, lty = 3) +
     geom_hline(yintercept = seq(2, 22, 2),
25              colour = "black", size = 0.4, alpha = 0.2, lty = 3) +
     geom_hline(yintercept = 14,
                colour = "black", size = 0.4) +
     # scale
     scale_x_continuous(breaks = c(1925, seq(1930, 1990, 10), 1999),
30                      labels = c(1925, "'30", "'40", "'50",
                                   "'60", "'70", "'80", "'90", 1999),
                        expand = c(0.005, 0)) +
     scale_fill_manual(values = cpal_qual_5) +
     scale_alpha_discrete(range = c(alpha_min, 1)) +
35   # coord
     coord_fixed(ratio = 5) +
     # theme
     ggtheme_min(base_size = font_size, base_family = font_family, grid = "n")↙
         ↳  +
     theme(legend.position = "none")
40
   # Plot qual-seq legend ---------------------------------------

   # sequence of alpha values based on minimum alpha and number of steps
   seqnce <- seq(alpha_min, 1, length.out = 4)
45 # alpha blended base colours
```

```
    alpha_blend <- sapply(cpal_qual_5, function (.x) AlphaRGBToRGB(.x, seqnce))
    alpha_blend <- t(alpha_blend)
    # transform into long format with positions and ids for ggplotting
    xy_coord <- expand.grid(x = 1:nrow(alpha_blend),
50                          y = 1:ncol(alpha_blend))
    xy_coord$col <- as.vector(alpha_blend)
    xy_coord$id <- seq_along(xy_coord[,1])

    # plot qualitative sequential legend
55  plot_qual_seq_lgnd <-
      ggplot(xy_coord) +
      geom_tile(aes(x = as.factor(x), y = as.factor(y), fill = as.factor(id)),
                colour = "white", lwd = 1) +
      scale_fill_manual(values = xy_coord$col) +
60    scale_x_discrete(labels = levels(counts_5$COD)) +
      scale_y_discrete(labels = levels(cut_interval(counts_5_mode$px, length = ↵
        ↳ 0.2))) +
      coord_fixed(4) +
      ggtheme_min(base_size = font_size, base_family = font_family, grid = "n")↵
        ↳  +
      theme(legend.position = "none",
65        axis.text.x = element_text(hjust = 1, angle = 90),
          axis.text.y = element_text(hjust = 0.5, angle = 90),
          axis.title = element_blank(),
          axis.ticks = element_blank())


70  # Merge legend and plot ----------------------------------------

    # a hack to add the legend to the plot and save the result
    AddLegQualSeq <- function (.plot, .legend, .path, .width, .height) {
      pdf(.path, width = 0.4*.width, height = 0.4*.height, useDingbats = FALSE)
75    grid.newpage()
      print(.plot, vp = viewport(x = 0.5, y = 0.5))
      print(.legend, vp = viewport(x = 0.82, y = 0.5, height = 0.6))
      dev.off()
    }
80
    AddLegQualSeq(plot_qual_seq, plot_qual_seq_lgnd, "./fig/plot-qual_seq.pdf",↵
      ↳  20, 15)
```

```r
    # Plot agewise-area chart -------------------------------------

    plot_agewise_area <-
      ggplot(filter(counts_5, Sex == "total")) +
 5    # main
      geom_area(aes(x = Year, y = px, fill = COD)) +
      # annotate
      geom_vline(xintercept = seq(1930, 1999, 10),
                 colour = "black", size = 0.4, alpha = 0.2, lty = 3) +
10    # scale
      scale_x_continuous(breaks = c(1925, seq(1930, 1990, 10), 1999),
                         labels = c(1925, "'30", "'40", "'50",
                                     "'60", "'70", "'80", "'90", 1999),
                         expand = c(0.005, 0)) +
15    scale_fill_brewer("", palette = "Set3") +
      # coord
      coord_fixed(ratio = 5) +
      # facet
      facet_grid(Age~. , as.table = FALSE) +
20    # theme
      ggtheme_min(base_size = font_size, base_family = font_family, grid = "n")↙
        ↳  +
      theme(strip.text.y = element_text(angle = 0, hjust = 1),
            panel.margin = unit(0, "cm"),
            axis.text.y = element_blank())
25
    ExportPDF(plot_agewise_area, "./fig/plot-agewise_area.pdf", 17, 17)
```

```
   # Plot small multiples -----------------------------------------

   # subset to modal cause of death for each year, sex, age
   counts_10 %>%
5    group_by(Year, Sex, Age) %>%
     filter(px == max(px)) %>% ungroup -> counts_10_mode

   # annotation data
   annot_dat <- data_frame(Year = 1930, Age = factor("30-34"), COD = factor("↙
     ↳ Infections"))
10
   # plot small multiples of all COD
   plot_small_multiples <-
     ggplot(filter(counts_10, Sex == "total")) +
     # aestetic
15   geom_tile(aes(x = Year, y = Age,
                   fill = cut_interval(px, length = 0.1))) +
     geom_tile(data = filter(counts_10_mode, Sex == "total"),
               aes(x = Year, y = Age),
               fill = "transparent", colour = "black") +
20   # annotate
     geom_point(data = annot_dat, aes(x = Year, y = Age), shape = 21, colour =↙
       ↳ "white") +
     geom_text(data = annot_dat, aes(x = Year, y = Age), label = "italic(A)",
               family = font_family, size = 4, colour = "white",
               hjust = -0.3, vjust = 1, parse = TRUE) +
25   # scale
     scale_x_continuous(breaks = c(1925, seq(1950, 1990, 20)),
                        labels = c(1925, "'50", "'70", "'90"),
                        expand = c(0.005, 0)) +
     scale_fill_brewer(name = "px", type = "seq", palette = "YlGnBu",
30                     guide = guide_legend(reverse = TRUE)) +
     # coord
     coord_fixed(ratio = 5) +
     # facet
     facet_wrap(~ COD, ncol = 5, as.table = FALSE) +
35   # theme
     ggtheme_min(base_size = font_size, base_family = font_family, grid = "n")

   ExportPDF(plot_small_multiples, "./fig/plot-small_multiples.pdf", 30, 20)
```