Basic principles for visualizing data in charts and presenting statistical formulas in print

(new version)

Opracowanie **Dariusz K. Chojecki**

Przekład **Karen Sayce**

Korekta **Marta Walkowiak**

Introduction

Presenting data in a visual manner using charts is often used in historical research that relies on mass sources or statistics that have already been compiled. Visualization serves as a valuable accompaniment to the text's content, and serves as proof that the content is verifiable. It also helps authors convince their readers of their point of view, and, above all, attract their attention by illustrating a particular issue, e.g., its structure, progression over time or certain co-dependencies. In other words, graphics appeal to human senses more effectively than a separate table or simply text. Despite the importance of visualization, the vast majority of charts, at least at their initial stage, resemble the famous "ugly duckling" of Hans Christian Andersen's fairy tale.

The main objective of this mini-guide is to identify those parts of a chart that can be altered to make it look like a small work of art, be aesthetically pleasing, attract the reader's attention and allow them to appreciate it more. This then allows us to convey the message that is most important to us. Following Cole Nussbauemer Knaflic, all the graphics featured here have been created using Excel, which is good for producing "ugly ducklings," but can also be used to turn charts into "beautiful swans" or at least resemble them half-way (we will not be discussing other programs here—if you are interested in using advanced tools, try **R** and **Tableau**, among others). The message is obvious—before sending texts and charts to the editor, it is worth working on the graphics first and avoiding the default settings since the program, inevitably, has no idea what we have in mind. This also involves choosing the right chart, i.e., the one most appropriate for our data and what we want to convey. As the author of this mini-guide, I have also come to realize that the form should not overwhelm the content, but instead benefit it.

As a rule, in academic papers, charts have no function without tables and the narrative content. In order to create these charts, we first need data, which is usually shown in the form of a table. This does not always appear in the actual text or appendix, and sometimes becomes the silent protagonist of our graphic story, as of course we need the spoken word or the text to present our subject. These factors, especially the former, will not, however, be the focus of our attention. This is because no recommendations will address how to use charts when we are presenting our research results to an audience at seminars, conferences, lectures, etc., which should be characterized by a certain dynamism and story structure. Instead, we will be focusing on the static components of charts, i.e., those that form a closed, formed picture of the issue that is embedded in the academic text. Whereas a speaker can build up their story gradually during the lecture before arriving at the final graphic, the author of a written text has to explain the idea of the issue without raising the tension or highlighting individual items on separate slides. This is not only because the chart, as a rule, should have an exploratory (evidential, explanatory) character—i.e., when the researcher has already finished looking for certain patterns—but also due to the sheer number of restrictions set by editors or publishers on the number of graphics possible and the size of the work. As well as the financial aspect, an important role is played by the retention of certain proportions between the actual text and the other parts of the paper. This problem is slowly becoming less of a concern now that various types of research data can be uploaded to online repositories designed for that purpose. In any case, we will treat the forms of the charts discussed here as independent parts of an academic paper which should also be communicable to readers with no prior knowledge of the underlying textual narrative. This leads to the rather trivial conclusion (which is not always respected) that charts should be given titles that answer the following questions: *Who?* or *What?*, *Where?*, *When?*, and also *Based on what?* The first three questions relate to the fixed attributes of the population under study. From them, for example, it is possible to uniquely identify the population we are studying, due to which it "becomes" unique, and one-of-a-kind. These attributes are possessed by all the individuals that comprise the collective, while they differ in the variable characteristic(s) (i.e., variables), which should be put into the title after the word *by*. Of course, chart titles show the evolution of

issues over time, or the interdependencies will already have a slightly different structure. Nevertheless, they should also include the constant features mentioned in order to unambiguously identify the subject of the study. The latter, especially in historical research, is also aided by indicating the source used to compile the chart. In this case, its notation may vary depending on whether we are dealing with archives, surveys (rare), official statistics or research papers, etc. However, this does not excuse us from having to provide the source of the data in a way that the intended reader can easily identify and, if necessary, can access it and its contents (unless the data is covered by a special confidentiality clause).

Choosing a chart

Now that we know what set of people, facts or events we are going to represent in our charts and what data we are going to use to accomplish this, our next step—assuming we have a clear research objective—should be to choose the appropriate chart. Despite popular belief, this skill is not commonly taught. At times, authors create charts without any deeper reflection on their visual aspect, or on their validity. There are dozens of types of charts, the choice of which can indeed be somewhat overwhelming for the creator. Therefore, we should highlight those charts that are most commonly used in visualizing the vast majority of data—these include column and bar charts (and their variations for cumulative values), waterfall charts, histograms, line charts, slopegraphs, scatter plots (and their weighted variation, the bubble chart), and box plots. To this, we should also add a chart called a treemap. We already have twelve, so let us limit ourselves to these in this mini-guide. It is not hard to spot that there is a notable absence here of the most commonly used graphic, the pie chart, whose use, without going into details, is not recommended, because its "attractive" form significantly impairs our perception of the data shown in it. At the end of this paper there are examples of charts based on selected "ordinary" data from the 19th and early 20th centuries relating to the town of Trzcianka in the northern part of Greater Poland. Sometimes variants of the same type of chart are also shown. When visualizing the data, a specific color scheme has been used; however, please note that in *Przeszłości Demograficznej Polski / Poland's Demographic Past*, we recommend that authors submit black-and-white graphics for printing, and use gray tones.

We most commonly work with data that illustrates the underlying structure or evolution of a given issue over time. In simple terms, the former means we have certain parts that make up the whole of the population. These can be categories of a qualitative variable, such as gender, marital status, household type, etc., or variants of a quantitative variable, such as number of children, age, time when an event occurred, etc. All absolute or relative frequencies (e.g., percent, per mille) of the occurrence of each category or variant should, when added up, give a total, overall count, or relative values, e.g., 100 or 1,000. If we have structural information, we can use a column chart (figs. 1–3), bar chart (figs. 4–5), slopegraph (figs. 10–11), or treemap (fig. 15) to visualize the categorical data. This order is not accidental, i.e., the first is the most commonly used, as people are best able to distinguish graphic features when they compare the features' heights. A column chart has a vertical orientation, while a bar chart has a horizontal orientation, and this is how they fundamentally differ. Although our perception of the latter is somewhat lower, it has one very important advantage in that it allows us to put longer labels next to individual categories, which are then easier to read due to the text's horizontal orientation, which looks more natural to the eye.

It is often impossible to keep the labels horizontal in column charts, especially when there are many categories or if their descriptions are longer. In this situation, the program will automatically apply diagonal text, or the author can change it to vertical, which will make the labels harder to read. This raises the dilemma of which type of chart to choose. Pragmatics dictates that we should use a column chart, as long as the text descriptor for the category under the x-axis is easy for the reader to take in and does not pose a cognitive burden. This recommendation also applies to continuous data (e.g., time, including age, distance, area) or quasi-continuous data (e.g., population, taxes, income), in other words, data that has an infinite or very large number of possible variants and therefore requires a *from–to* interval notation. In older versions of Excel, the histogram was unavailable. Therefore, users had to resort to a column chart, make the appropriate grouping and enter the intervals as plain text. In newer versions the histogram is included, and its labels have formal interval notation. The user thus no longer has to do the grouping, i.e., determine how often each variant occurs, and individual data will suffice. It should be noted that

displaying continuous data requires that the columns corresponding to their levels be adjacent to each other. While in a histogram this is taken into account "automatically," in a column chart this is no longer the case (there is a tacit assumption that the bulk of the data will be categorical). To solve this problem, we should set the width of the gap to zero. Why do I mention this? Because the histogram gives no option to display the individual components for each *from–to* variant. In this situation, we need to use a special type of column chart, as we shall see below.

We are often interested in comparing not only the categories themselves in terms of values, but also their components, figuratively called "elements" here. Let's say we have data on gender and the main causes of death. In this case, we can juxtapose each element's value side by side for a given category, or use a cumulative chart to show "concisely" either the absolute or relative incidence (e.g., expressed as a percentage) of the main causes of death, which are the elements here. What we obtain are two columns divided into their corresponding parts. By comparing them, we can form an initial idea of structural similarities and differences, to answer the question of whether gender affects the proportions in our topic of research in some way. From there, it is just a short step to examining any correlations.

When comparing—most commonly, two structures—we can successfully use a slopegraph, which is a kind of variation on a line graph for categorical data (figs. 10–11). In this case, the components, i.e., category elements, are replaced by markers that represent a specific value of the share. We can also look at the structure of an issue (e.g., attitudes toward the use of contraception) in terms of whether it was formed before or after the occurrence of a particular fact, event or action. Markers for elements that have the same meaning are connected by a straight line, the angle of which allows us to automatically assess the direction and strength of the changes (decline, stagnation or stability, growth). However, there is a limitation to this type of chart—it becomes less intelligible when a high proportion of the elements being compared have significant dissimilarities in terms of value level formation and there are numerous line intersections. This problem becomes more of an issue if the number of elements being compared is large.

To show structural issues we can also use a treemap, especially when we plan to illustrate multiple categories that also contain elements, the number of which can also be significant (fig. 15). Data on the number of birthplaces by top-level and lower-level government administrative units comes to mind here, as the main categories could be parishes or districts, and the elements could be villages. The name of this type of chart is somewhat deceptive. It is in fact a rectangle divided into smaller rectangles, which can then be divided into even smaller rectangles. Their area corresponds to the value of a given category and the elements. The algorithm for creating the chart allows us to get an idea of our data's hierarchy. Starting from the top left-hand corner and looking down to the bottom right, from the highest to the lowest count, we can see the categories (in relation to the total) and their elements (in relation to a given category). This appealing design makes it possible, for instance, to observe how the proportions of element values within each category are distributed. In contrast, when we compare elements of equal importance but operating in different categories, we may come up against difficulties in assessing their contribution to the whole. This is because rectangles of the same or similar area often have different shapes, which is obviously related to the attempt to integrate them into the whole. This type of data representation can also be used for comparisons over time, such as between two moments or periods in time (two separate charts juxtaposed one below the other).

To compare the development of trends over time, and the dynamics, a line graph is primarily used, on which more than one series of data can also be shown, such as the evolution of vital event elements (figs. 7–9). When using a line graph, we must always keep in mind that the units of time—months, quarters, semesters, years, decades, etc., —should be the same. This is because mixing them together in a time series distorts the underlying trend. This is also the case when the intervals or moments in time are not the same, e.g., 1686, 1687, 1688, 1701, 1705, (years are skipped) and the user does not select the *date axis* option in the *axis type*, which would allow the influence of time on the issue's evolution to be taken into account. What often happens is that, when this is applied, sharp increases or decreases disappear, as if by magic. These can also be the result of territorial changes in the administrative units (scale-ups or subdivisions) for which the data is aggregated. This too should be kept in mind when illustrating the evolution

of an issue over time. The ideal situation is when we have complete data for homogeneous time units and administrative units that do not change their area. If this is not the case, it is worth considering whether the line graph we are about to produce will accurately portray the issue, whether there will be too much distortion and, thus, whether our depiction of the trend will be deceptive. However, if we do decide on this type of graph, then it must be accompanied by the appropriate commentary. Approximation is another problem related to time units, i.e., in order for the program to interpret these units correctly, they must be saved in either numerical or date format, so that in the *axis type* we can select the *date axis* option. Unfortunately, when we have a textual record of time intervals in the form of *from–to*, which is very common in historical demography, the program is unable to recognize their meaning. In other words, with non-uniform spans, e.g., 1831–1835 (5 years), 1836–1840 (5 years), 1841–1850 (10 years), the Excel chart will not graphically reflect the impact of the longer interval on the development of the issue being analyzed. A kind of workaround for this problem is to use numerical values determined according to the midpoints of the intervals, i.e., 1833, 1838 and 1845. These averaged values can now be joined by lines; this procedure, however, requires appropriate commentary.

Also worth mentioning is that line graphs can be used to illustrate the seasonality of events. For seasonality indicators, where calculating the average level of the events being analyzed is often taken as 100, it is advisable to set the intersection with the horizontal axis in the *axis options* to this value, so that we obtain a reference line. This is formed by the time units axis. What is above it has higher deviations from the average, while what is below it has lower deviations. This graph can easily be transformed into a layered one, but we must not forget to use the right kind of transparency in the color fill when two series of data whose fill contours intersect. This allows us to see the course of the previously covered data series.

We often need to check whether there is any correlation between the individual values of two quantitative variables. To do this, we can successfully use a scatter plot, which shows the data markers in a Cartesian coordinate system (figs. 12–13). Based on the arrangement of multiple points (streaks), we can tentatively

infer the existence or absence of a relationship between the variables (e.g., a summary of the proportion of the Polish-speaking population and the rate of infant deaths in selected districts). We can also make decisions about performing data transformation and, most importantly, selecting an appropriate measure of correlation. If we are analyzing multiple variables, they are shown on separate mini-charts in a panel arrangement. With this type of chart, we need to watch out for different data with the same coordinates, e.g., a point $x = 10$ and $y = 10$. In this situation, one point will cover the other on the graph. The overall picture of our issue will therefore be less clear. To counteract this, we can apply weightings to each point, which will show how often the overlapping points occur, e.g., we may find that most of the points receive a weighting of "1," some "2," and others, probably less numerous, receive a weighting of "3," etc. Accounting for this involves adding a third column to the data, which will give the number of occurrences. This will be used to scale the size of the point markers with the size of the circles. Obviously, the larger the scale, the larger the area of the objects. The bubble chart, which is a variation on the scatter plot, can also be used when we want to give a weighting to a subject that is not determined by the number of occurrences. I mentioned earlier a sample comparison of the proportion of the Polish-speaking population and the rate of infant deaths in selected districts. As is generally known, districts are not the same in terms of population or demographic events. To account for this, we can use weightings in the form of the number of live births or of infant deaths, or others that are reasonable to use. This will provide the reader with important additional information, e.g., it will be more relevant to assess the issue for the metropolitan borough of Poznań, which is the most populous and thus with the highest absolute number of births and deaths in the group of administrative units studied, and as signaled by the bubble size assigned to it. Two aspects are worth noting here: firstly, that the data for the weightings should be sorted from largest to smallest, and secondly, the areas of the circles on the graph should have a visible outline—this way we can avoid covering one object with another or overlapping them.

Both the scatter plot and the histogram are created from individual statistical data, i.e., information that, as a rule, is not the result of grouping but can be formed by aggregation, such as the number of births, deaths, marriages, or the

state of the population in individual parishes. Here the statistical units are administrative units, not individual people, households or families that are most often the subject of historical demographic research. Among the charts based on individual data, the box plot is of particular importance in addition to those mentioned above. It can be used for exploratory data analysis (EDA), such as the age of individuals, or to check or compare the distribution of a continuous or quasi-continuous variable, including spotting outliers relative to others (fig. 14). Understanding how this works is far from intuitive, as it requires some in-depth knowledge. The cornerstone of this statistical chart, which is a relatively recent addition to Excel, are positional averages, i.e., the first, second (median) and third quartiles. On the chart they form a box, the width of which is irrelevant. The lower edge of the box serves as the first quartile value, the upper edge is the third quartile, while the dividing line is the median. The center of the box contains half of the observations, while below and above it are a quarter each. The edges of the box generally form the minimum and maximum values, but this is not always so. The height of the box is the difference between the third and first quartiles. As suggested by the chart's inventor, if any value is one and a half times the height of the box from its lower or upper edge, then it is considered an outlier and marked with a circle. In this situation, the outliers are positioned at the level of the last value in the dataset that does not exceed the critical value of one and a half box heights. The chart can also display the symbol for the arithmetic mean, which further allows for an initial assessment of the distribution's skewness.

Our last basic chart is the waterfall chart, which can also be successfully used to visualize demographic data, especially the population balance (fig. 6). In other words, it allows you to trace the population stock of a given population and the level of demographic events affecting them, beyond the initial one, of course. As a reminder, demographic events, i.e., births (+), deaths (−), inflows (+) and outflows (−), are components of the balance of the population. Their numerical extent is shown on the waterfall chart by means of rectangles of the appropriate height, while the population stock should take the shape of columns "seated" on the horizontal axis of the category on the chart. And in order for their height to be shown from zero, we need to select the relevant rectangle symbolizing the population stock, and then in the *data series* options we check the box *set as*

sum. From the balance of population equation, it follows that the population stock at time *t* is the sum of the value of the population stock at time *t–1* and the components between the two moments in time. If our data is perfectly accurate, then the last component, e.g., the outflow, will be linked to the adjacent population stock by a horizontal line. In practice, we will see a discrepancy between the delineated population stock based on the balance of the population (using the ongoing registration of demographic events) and the population stock obtained from the census. We can also add this discrepancy to the waterfall chart as the last component, i.e., the connecting line between it and the adjacent resulting population stock will then have a horizontal orientation. In historical demography, it is rare to have data from current-day population registration, so as a rule, the migration balance is estimated on the basis of vital statistics and population stocks. The result of these estimates can also be transferred to a waterfall chart.

Chart design

We usually think of aesthetics in terms of beauty that is hard to define, a certain structure and the right proportions, along with the way we convey our ideas to make the viewer feel the way we intended, which in our case is positive, i.e., not making them feel uncomfortable because it is too hard, impossible even, for them to grasp content that has superfluous elements or an inappropriate format. As can be noticed, when creating charts, science meets art, i.e., the designer sends a message to the reader mainly based on their perception of shapes—pre-defined, as it were. According to the principles of *gestaltism*, our mind distinguishes objects based on proximity, similarity, spatial confinement, enclosure, continuity, and connection. Importantly, we do not have to be aware of this at all in order to correctly interpret visual graphics, which usually contain markers, lines, rectangles or other areas of a certain size, shape or fill in color, less often in grayscale. However, the onus is on the designer to make optimal use of these features in presenting their content to their readers. Rarely do we show just one or two values using charts. As a rule, there are also many categories (variants) or series of data. These categories function in a specific space and are a benchmark for themselves. Before we give a brief outline, or rather point out some principles that are worth respecting to ensure a more effective understanding of the content, let us focus on their setting and try to

answer a few questions, keeping in mind that the answers are only a certain suggestion to consider when tailoring charts to our needs.

What shape should the area of the chart be, i.e., where the elements illustrating the values are located? Unless we are dealing with a scatter plot, which generally requires a square, the vast majority of graphics will look aesthetically pleasing when using a length-to-height ratio of 5:3 in a horizontal orientation (e.g., a column chart). These ratios are close to the golden number that the ancient Greeks used. We only have to look at the screen of a modern TV or computer to see that this kind of approach is also well known to modern designers. Therefore it would be a good idea to apply this kind of proportion to a bar chart. However, there are times when the chart area has a large number of categories and needs to be stretched vertically. In this situation, it is advisable to ensure that the value of the ratio in question is not greater than 5 in the divisor.

What fill should the chart area have? White. Any other color in a research paper is unnecessary, contributes nothing, and is simply too much effort for the reader. The argument in favor of a white background is that the main text of an academic publication is set on this background, and very rarely on a gray one, which is sometimes used for "excerpts," i.e., commentaries, explanations, appendices, and small extracts from textbooks. The underlying reason for this is that on a white background almost all colors and shapes are clearly visible and recognizable. Contrast is important here; in other words, black dots on a white background will attract our reader's attention faster than on a gray background.

Which elements of the chart may be unnecessary? Quite simply, everything that is irrelevant to perceiving the content or unessential for an accurate and quick interpretation of the chart's message should be removed or placed at a lower level in the visual hierarchy. So we can use the principles of *gestaltism* to reduce unnecessary elements if we think that this will not adversely affect how the chart is perceived. This usually applies to various types of borders, i.e., once they have been removed it is obvious what is important and the reader can better focus on the relevant content. Often the chart is overloaded with grid lines, including auxiliary ones. It is therefore recommended to use these elements sparingly or not at all since, especially with line graphs, they are a certain visual burden and disturb the reader's perception of a given trend. If you decide to leave the grid

lines, give them a light appearance by using gray tones, low thickness (hairline) or dotted lines. This will help bring the size of the issue, its development over time, structure, interdependencies, etc., to the fore, in keeping with the type of data being visualized.

To show the magnitude of our subject "more precisely," labeling can be used. However, be aware that too many values reduce a reader's perception, overwhelming them with details. Therefore, it is worth taking the time to choose the most appropriate labels which you feel will best draw attention to what is most important in the content you are conveying. Obviously, labels with values should not significantly interfere with how the graphic elements showing the magnitude of the issue are perceived, so you cannot always count on the software's placement options. What comes into play next is placing them manually, which in terms of editorial design should be characterized by a certain consistency (i.e., changing the position but maintaining a similar proximity to the relevant marker). It is recommended that the labeling of columns or bars be inside these graphic objects as far as their layout allows. The chart does not need to show all the numerical details. These should be included in a table (the source of the chart), which can be included within the text, in an appendix or in a research data storage facility. Numerical labels inserted in the center of the chart elements can be made more lightweight if you use white for their values, as long as you allow for a fill that creates sufficient contrast. Bold fonts also work very well.

I am not advocating the elimination of axes that illustrate the scale of a given issue or are lines under which descriptions of categories, variants, or time units are placed. After all, they are a certain "plane" of reference that allows a quicker understanding of the size of the data being visualized, although it is important that in terms of form they should operate in the background. On the other hand, I am in favor of omitting thick supporting line markers on value axes, and opting for main line markers, orienting them outward, i.e., toward the values to which they refer. The latter, if possible, should be expressed in whole numbers, without unnecessary zeros after the decimal point, and separated in the case of numbers of four digits and above. The display of numbers on the value axis should be chosen so that about five of them appear (e.g. 0, 25, 50, 75, 100). Where did

this number of five elements come from? Well, our mind is able to automatically count items that do not exceed five—the number of fingers on one hand. This has been proven by an experiment with identical dice (when six dice are thrown from a cup, the participant is often unable to give the exact number of dice instantly, i.e., quick answers are given, e.g., five, six, or seven, with the proportion of incorrect answers being high).

Based on the above, we can conclude that reducing the cognitive load, and thus more effectively attracting our reader's attention, is possible when the number of elements analyzed is not too high. Of course, this raises the dilemma of whether it is worth reducing the amount of data to make it easier to understand. The answer is to seek some middle ground, make certain compromises and remember that excess details can be shown elsewhere. A type of redundancy is also the use of two x-axes, but this does not at all mean duplicating data, because this simply should not be done, but that two axes can represent two different scales for two different things, such as mortality and morbidity. This approach is tempting and often looks impressive, but nevertheless adds a certain cognitive load for the viewer, as they have to put more effort into distinguishing between data values. All European languages are read from left to right. It thus follows that the value scale for a given subject should be placed on the left side of the chart. Therefore, if there is a way to dispense with a scale on the right side, this is recommended, although it is not an absolute requirement. On the other hand, we should try to adhere to the principle of starting the axis from zero in the case of data on a quotient scale. Sometimes the opposite practice is associated with manipulation of data, as this can give the viewer the apparent impression of significant changes. Of course, it is not always absolutely necessary to apply this rule. This applies to data with low variability, with values "far" from zero, e.g., of the average life expectancy of newborns in regions or subregions of the same country in close cross-sections of time, shown by means of box plots. In any case, this deviation should be seen as confirming the rule that arises from the fact that data on a quotient scale—as the name suggests—must have a reference point in order to talk about proportions at all. When there is no reference point, we are dealing with data on an interval scale, which allows us to define only the difference or order, and therefore what, in the case of numbers, has less cognitive potency.

Making charts three-dimensional, i.e., giving them depth and therefore prominence, seems to be to the novice in data visualization (I include myself here) an attractive way to convey statistical information. Of course, what we have here is actually an imitation of three-dimensionality, which adds no substantive value to the reader's perception of the data, and, moreover, makes it difficult to approximate the level of the data's values. Simply put, using this form is unnecessary, redundant and is cognitive overload for the reader. Beauty lies in simplicity. Things are different when visualizing something that is indeed being shown in three dimensions, such as the number of infant deaths by calendar month, average temperature in degrees Celsius and life expectancy in days. In this situation, the use of a three-dimensional chart is justified since its depth is the result of the variable's specific values. In any case, this is a transformation of a three-dimensional space into a two-dimensional space, which naturally causes distortions, in other words, the inability to assess the interrelationships between the variables fairly accurately. It does, however, give us the chance to spot some dissimilarities in an issue's evolution under the influence of different variables and to make a very rough assessment of the strength of their interaction. True 3D charts are therefore more exploratory than explanatory. It is well worth looking for such alternatives, thanks to which three variables, not to mention a larger number that surpasses our perceptive capabilities, can be depicted in a classical 2D-type chart. For this purpose, for example, a bubble diagram can be used. It allows us to visualize numerical data in the context of the position and size of the markers in a Cartesian coordinate system. On the other hand, in a simple scatter plot the third variable can be a category, expressed in either a color or a shape. It follows that, by referring to position, size, color, shape and even style of border, it is possible to visualize more than two or even three variables on a two-dimensional plane. Unfortunately, the quality of the message is not (fully) satisfactory due to the heterogeneity of the way the data is shown. Therefore, such mixed solutions should be used with caution, or else we should use special charts; however, the rules for constructing these require advanced knowledge, i.e., understanding them is not intuitive.

What color and sequence of the graphic items symbolizing values should we use?
This question, of course, depends on the type of chart. The most common are

column and bar charts and line graphs, and we will focus on these. The basic rule is to use subdued colors, such as gray or blue tones. Opting for highly saturated colors is not recommended, nor is using their many different palettes or gradients, whose color variations within a category are meaningless. Vivid and/or mottled colors on the chart are tiring, garish, and unrefined (in my editorial work I have come across diagrams that in their preliminary version had canary-colored backgrounds and pink columns, never mind the other details). A certain kind of subtlety is important, as well as highlighting the elements we really want to draw the reader's attention to in more intense colors. However, the prominent elements in the visual message should not take up a lot of space, i.e., up to one-tenth of the area, in order for the reader to effectively focus their eyes. So the one important word here is moderation. This is an important part of the graphic strategy because, to reiterate, it allows us to highlight what we actually want to focus on. A white background is also our ally here as it allows us to "bring to light" the graphic elements that symbolize the values. If the latter dominate the background due to their prominence, size or labels, our message will be lost because from the very outset the reader's attention will be directed towards the chart's non-essential content.

An interesting approach is to use neutral white to mark the edges of histogram columns or treemaps. This gives increased brightness to the graphics—indeed, not introducing another color, such as black, helps reduce the cognitive load. We use what is called a general background. The latter is a kind of reference map for colors and shapes, i.e., for the backdrop. Thus, if we want to bring in "supporting actors," e.g., projected data alongside actual data, it should not be too conspicuous. Reducing the intensity of the color, the thickness of the lines, or changing the outline to a fine dot or dashed line will meet this goal.

Problems in perceiving content can arise when using a continuous color scale from a single-color range. This should be avoided and we should opt for progressive tones, with corresponding components that allow our vision to make the appropriate distinction. Without going into detail, if we use hues of a particular color we must take into account that we can only distinguish adjacent color variants when their total number typically does not exceed five (from relatively weak to relatively strong saturation). It is best to aim for relatively few

colors or hues, as this reduces the cognitive load and allows us to absorb the content of the visuals more quickly—instead of discouraging us, it encourages us to look carefully at the chart. In consideration of people with color vision deficiencies, it is worth considering doing away with red and green fills in our charts.

To reduce the number of colors or their hues, we can sequence the items in the legend, especially when we are visualizing components of individual categories or time series, as long as their number is no more than five, by convention. It is important to position the legend properly, i.e., either horizontally or vertically, depending on what is most beneficial. For charts showing the evolution of a given issue over time, using a legend to reduce the colors or shapes is harder if the lines that show dynamics intersect. In this case, it is better to use data series labels for selected markers, and also to make use of leading lines. Of course, the idea is not to label the series for each marker that the line passes through, but just for one "representative" of each series. The question will arise: for which one, though? For the one where the use of a label with a leading line will be the clearest, will stand out well on a white background, and will interfere as little as possible with the reader's perception of the other graphics. As can be seen, there is no ready recipe here, as everything depends on the layout of the data being visualized and the free white space available. It is here that we should place the description or legend content, (limited to a minimum), so that it does not dominate the foreground or cover other pictorial elements. Unfortunately, we are not always able to meet this condition.

The order of visual items that symbolize values on column and bar charts generally depends on whether the data represents a qualitative (categorical) variable or a quantitative (continuous, incremental) or quasi-quantitative (ordinal) variable. For the former we can be guided by the number of categories or their components, ordering them from largest to smallest. This step will allow us to obtain a more orderly picture of our subject of interest, eliminate or mitigate the adverse effect of different levels of data (i.e., the "jaggedness" of the picture) on its readability, and, above all, make it easier to grasp the essential and non-essential more quickly. So how do we carry out the data sorting? For bar charts, putting the numerically dominant category at the very top seems to

work well, given that we start reading from the top (sorting from smallest to largest). With column charts, on the other hand, placing the numerically dominant category on the left side of the relevant area of the chart is a good option, since in our Western European culture we read and write from left to right (sorting from largest to smallest). However, these recommendations can no longer be applied when visualizing values of quantitative or quasi-quantitative variables, since here the order is determined by the order of the variables themselves expressed in points or ranges, i.e., from the smallest to the largest of their variants (most often) or from the largest to the smallest (less often). Columns or bars should have a uniform color, such as gray; it is an artistic error to choose different colors for each of these graphic features. We should only play with color when we want to indicate some value(s) that play an important role in our narrative. To reiterate, highlighting only works if it is used sparingly. However, it is also effective if we use no more than two colors for accentuation, e.g., a basic one with a "background" function, such as gray (the vast majority of graphic items that symbolize values), and a highlighting one, such as orange or blue (the overwhelming minority). This graphical strategy is used by Cole Nussbauemer Knaflic, while the two attention-grabbing colors, interestingly enough, have for some time now been included in Excel's default chart template.

A word on formulas

The general principles for composing mathematical formulas are excellently described in *Jak pisać i redagować. Poradnik redaktora. Wzory tekstów użytkowych* (*How to Write and Edit: An Editor's Guide. Examples of Practical Texts*) by Ewa Wolanska et al. We will, however, focus here on the aspects of form which are most important when showing statistical (i.e., mathematical) formulas. In a journal they should be isolated from the text and put between paragraphs, with a legend under the formula or formulas explaining the symbols used, if their meaning is not commonly known. Mathematical formulas should be designed in the same format as the main text of the article, with the corresponding typeface, size, and placement of their components, which is also facilitated by the built-in equation editor in a given program. Sometimes we have to work with elaborate formulas. When carrying to the next line, the mathematical operator sign should be repeated in it (for multiplication, we recommend using the "×" sign). However, to ensure some visual tidiness, it is a

good idea, e.g., for strings containing equations and written over several lines, to center them relative to the equal sign. This is not compulsory when submitting an article proposal for initial evaluation by the editorial committee and reviewers, as it requires the use of more advanced formatting methods, part of which is also the use of appropriate indentation. In any case, the designs must be editable, i.e., they should not be an embedded image.

Each formula is marked with the corresponding consecutive number given in semicircular brackets without a period. This denotation should be in a specific place and at a specific distance from either the beginning or the end of the design so that it can be clearly distinguished from it. Also, the rule of thumb that it is placed against the right margin. When writing fixed values in formulas, do not use number separation. However, we should apply this option for larger values shown in charts and tables when separating five-digit numbers, e.g., 37 251. When a four-digit number is included in a group of five- or more digit data, it should also be separated, e.g., 2 541. Negative values are written with an en-dash preceding the number without a space. We can also use the en-dash to indicate periods or intervals: no spaces are added between it and the numbers. It is recommended that in formulas, non-whole numbers be expressed using decimal fractions. Of course, this point does not apply to division indicated with the fraction bar, which must be longer than the longest expression on or under it. A fractional expression enclosed in brackets should also be entirely contained within it. The same point applies to the root sign. Latin and Greek symbols are written in italics. With their multiples, we do not use spaces unless the notation consists of two or more characters and, as a rule, we omit the multiplication sign, which is also not required when we have a multiple of a number and a symbol. Finally, it is worth mentioning that in a formula set off to a separate line in which there is the summation symbol, i.e., sigma, we put the beginning and end markings of the summation above and below the sigma, respectively, while we put them to the right of the symbol at the top and bottom when, for some reason, we want to embed the formula in the text.

Concluding remarks

The recommendations outlined here do not close the entire catalog of issues that need to be addressed, as they only touch upon certain areas pertaining to the

importance of form when designing various types of tables and figures. The purpose here is to encourage the reader using in-built templates to modify them in order to clearly and precisely present the main idea behind the visualization of their own data. There are no ready-made solutions here but more general advice that the reader can incorporate. The most important of these, in terms of classic aesthetics, is that when designing charts beauty hides in simplicity, moderation and appropriate proportions. The more irrelevant items we manage to eliminate from the chart, the better located those items (and any descriptions) that represent the topic will be. And then the faster the reader can focus their attention on the essential message of the chart, the happier we should be with the results of our work. Before we display the results of our research, it is worth considering the different options for our graphics, both in terms of the choice of chart, the way the data will be aggregated, and the form of the presentation itself. If there is a significant degree of complexity in the data that illustrate our topic, and informational noise caused by its flow and multiplicity, then we should be bold and decide to turn one graph into many—and shown in a single panel.

Source Literature

Biecek, Przemysław. *Odkrywać: Ujawniać: Objaśniać: Zbiór esejów o sztuce prezentowania danych*. 2nd ed. Warszawa: Fundacja Naukowa Smarter Poland.pl, 2016.

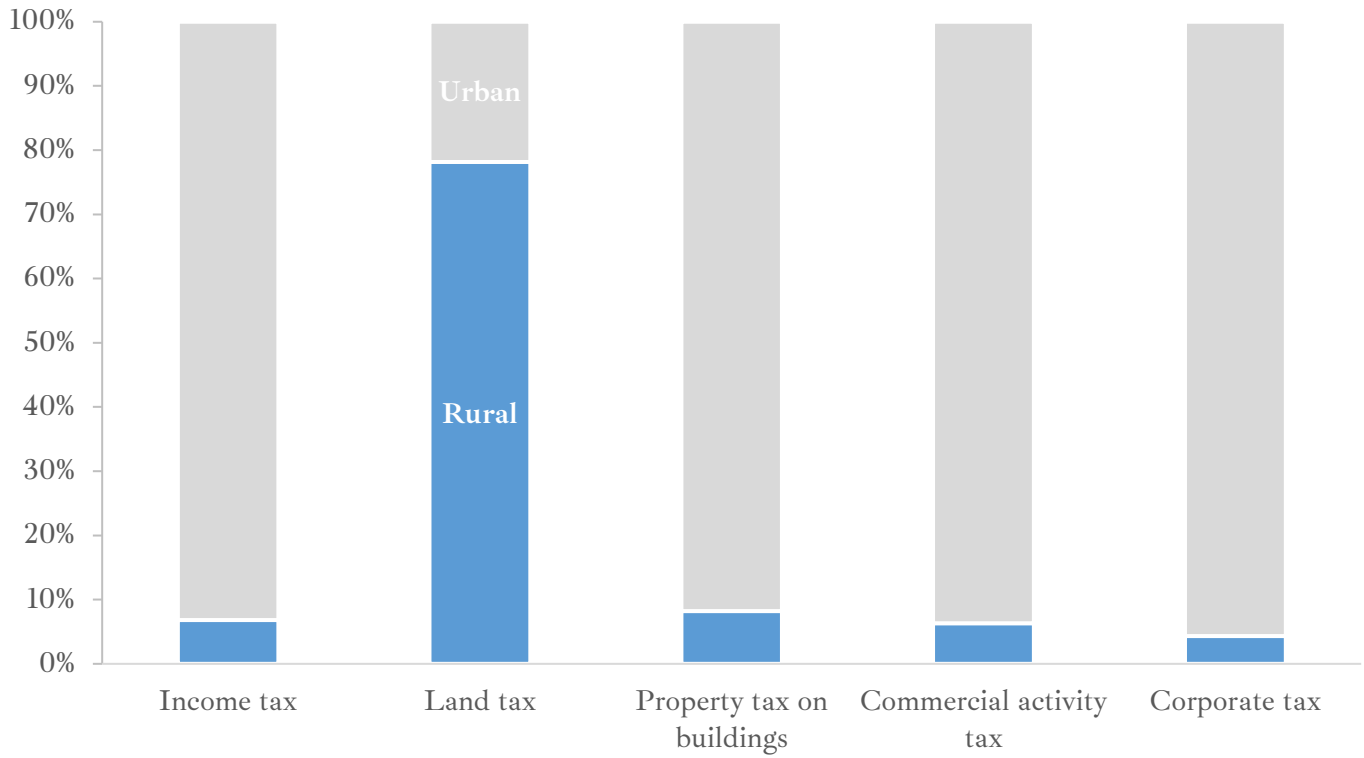
Knaflic, Cole Nussbaumer. *Storytelling danych: poradnik wizualizacji danych dla profesjonalistów*. Gliwice: Helion, 2019.

Wilke, Claus. *Podstawy wizualizacji danych*. Gliwice: Helion, 2020.

Wolańska, Ewa, Adam Wolański, Monika Zaśko-Zielińska, Anna Majewska-Tworek, Tomasz Piekot, eds. *Jak pisać i redagować: poradnik redaktora, wzory tekstów użytkowych*. 2nd ed. Warszawa: Wydawnictwo Naukowe PWN, 2022.

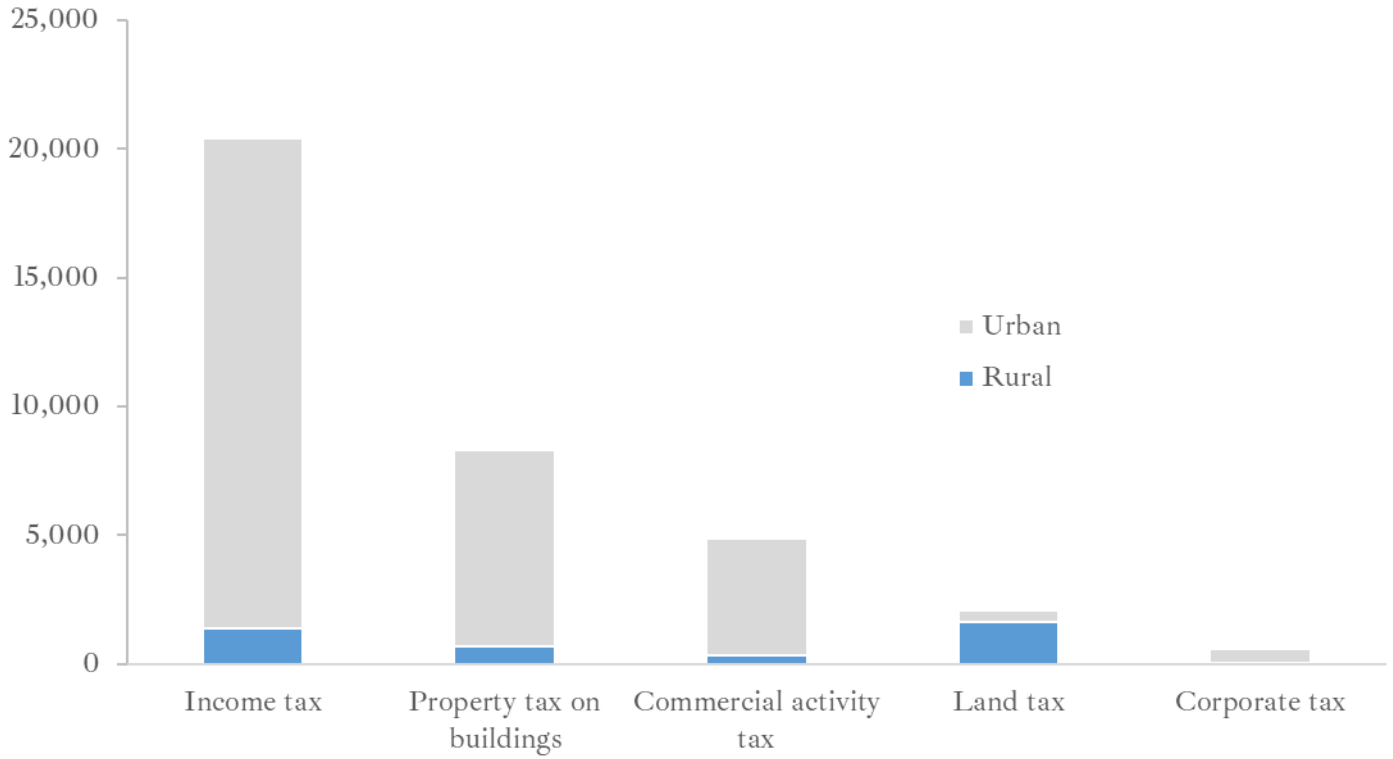
Appendix (examples of charts)

Figure 1. Relative share of amount of tax in Trzcianka's urban and rural areas in 1905, by each tax category



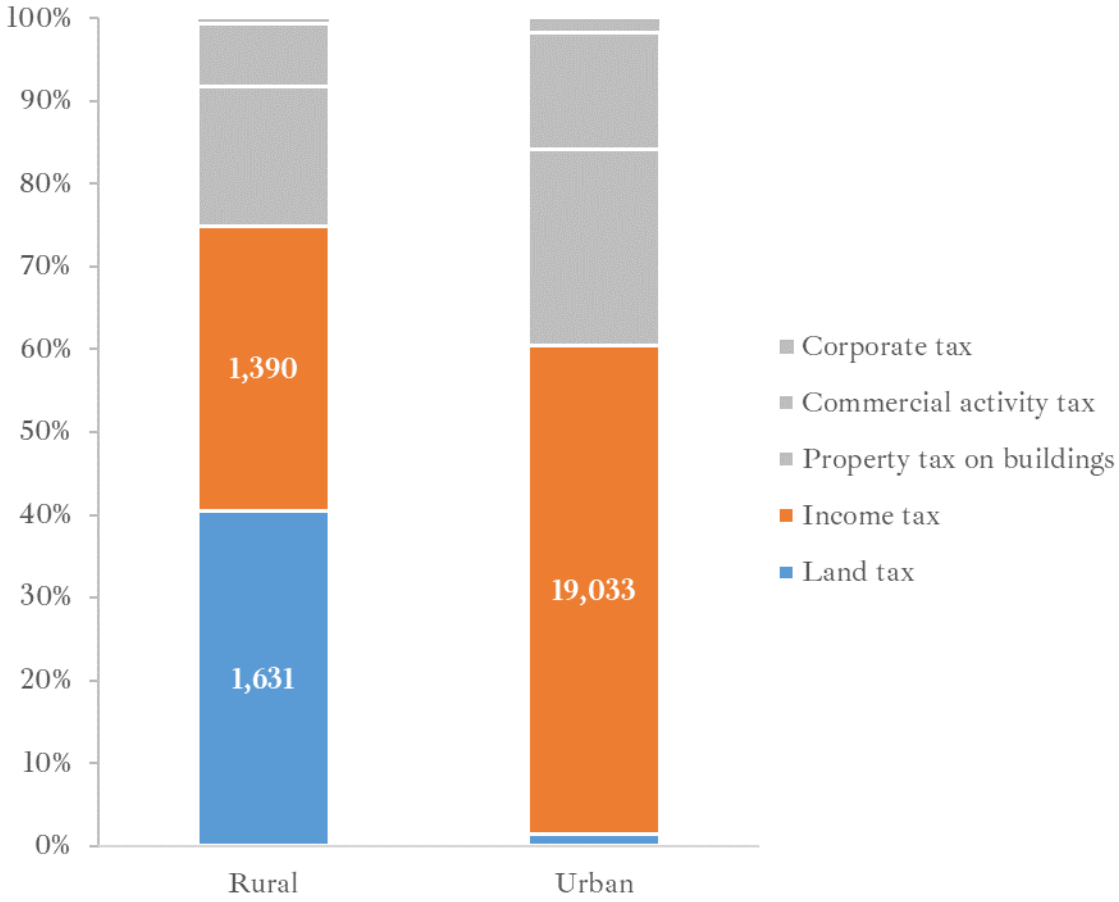
Source: own work based on Geheimes Staatsarchiv Preussischer Kulturbesitz, I HA. Rep. 77 Ministerium des Innern, Tit. 2686 Nr. 5.

Figure 2. Absolute share of amount of tax in Trzcianka's urban and rural areas in 1905, by each tax category



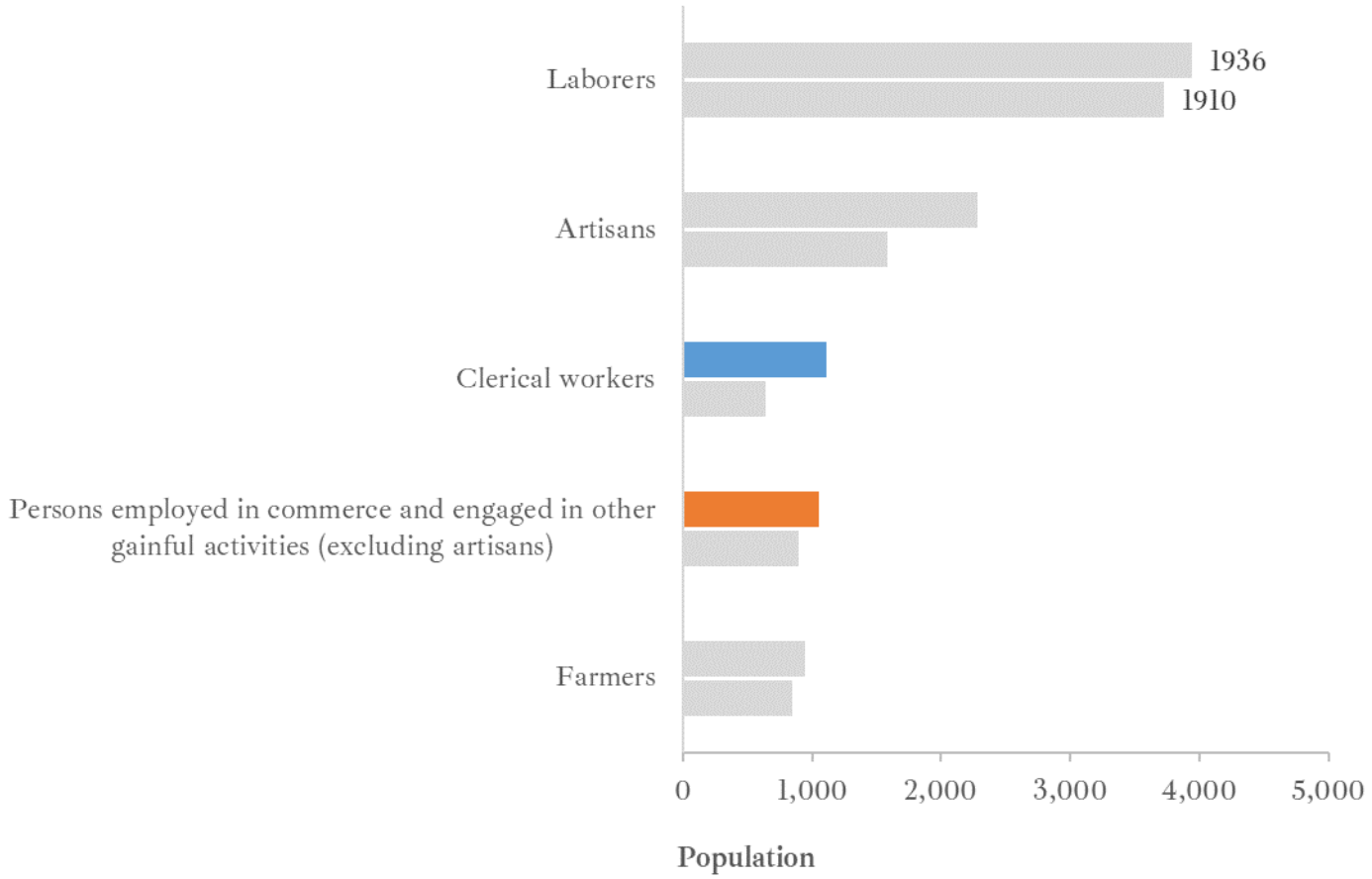
Source: own work based on Geheimes Staatsarchiv Preussischer Kulturbesitz, I HA. Rep. 77 Ministerium des Innern, Tit. 2686 Nr. 5.

Figure 3. Relative share of amount of tax (in marks) in each category in 1905, divided into urban and rural Trzcianka



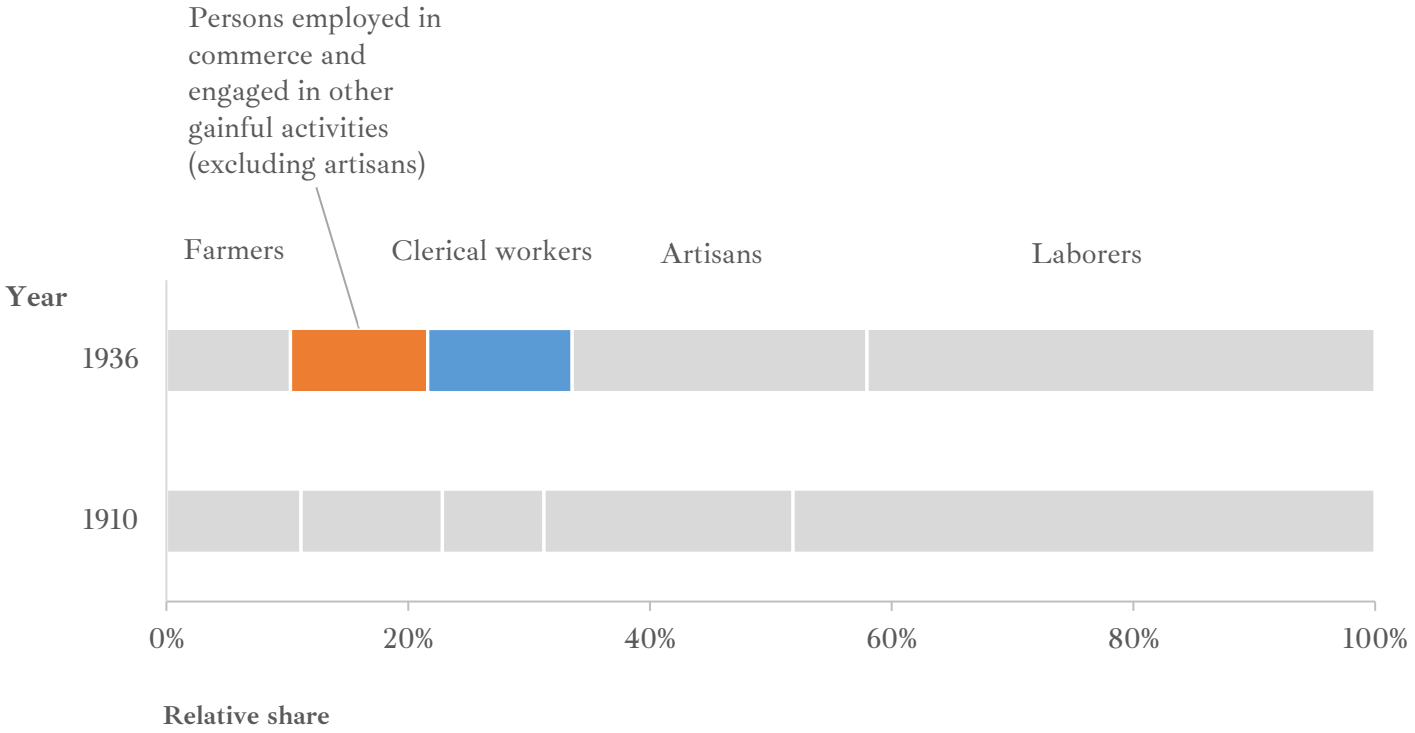
Source: own work based on Geheimes Staatsarchiv Preussischer Kulturbesitz, I HA. Rep. 77 Ministerium des Innern, Tit. 2686 Nr. 5.

Figure 4. Population of the town of Trzcianka in 1910 and 1936, by source of income



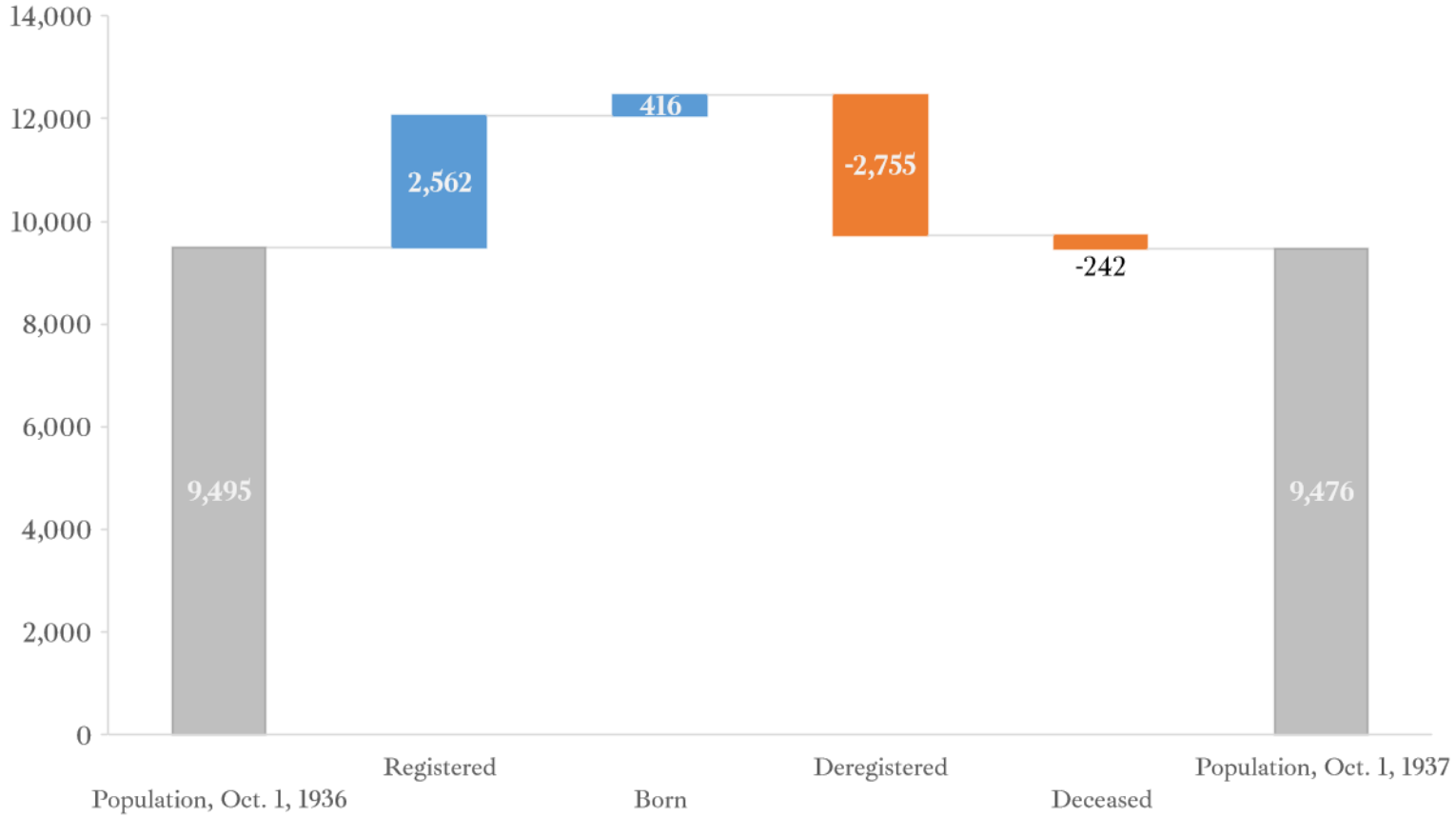
Source: own work based on Archiwum Państwowe w Poznaniu Oddział w Pile, Akta miasta Trzcianka, sygn. 55/15/0/9/300.

Figure 5. Population of the town of Trzcianka in 1910 and 1936, by source of income



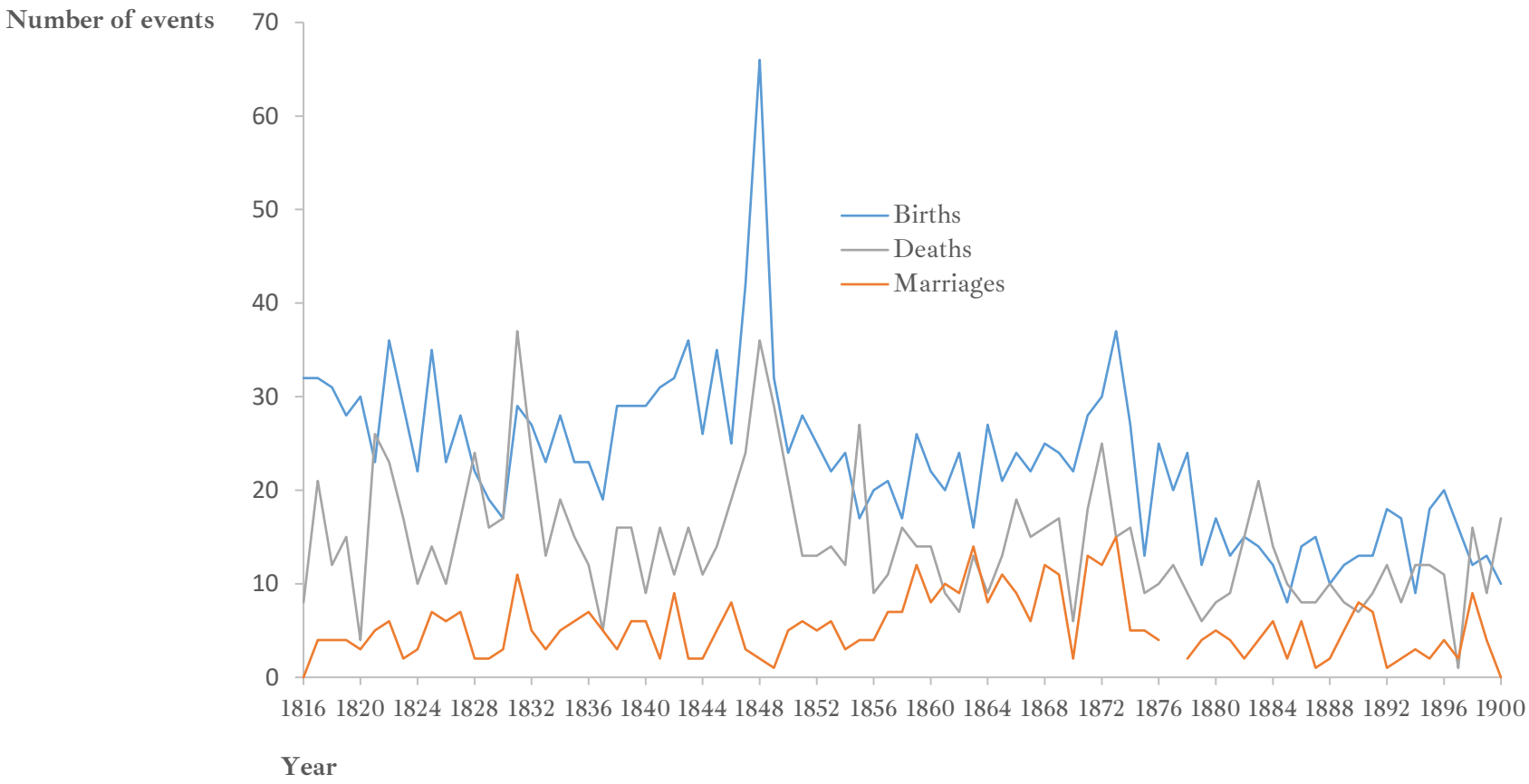
Source: own work based on Archiwum Państwowe w Poznaniu Oddział w Pile, Akta miasta Trzcianka, sygn. 55/15/0/9/300.

Figure 6. Components of the population balance for the town of Trzcianka from October 1, 1936 to October 1, 1937



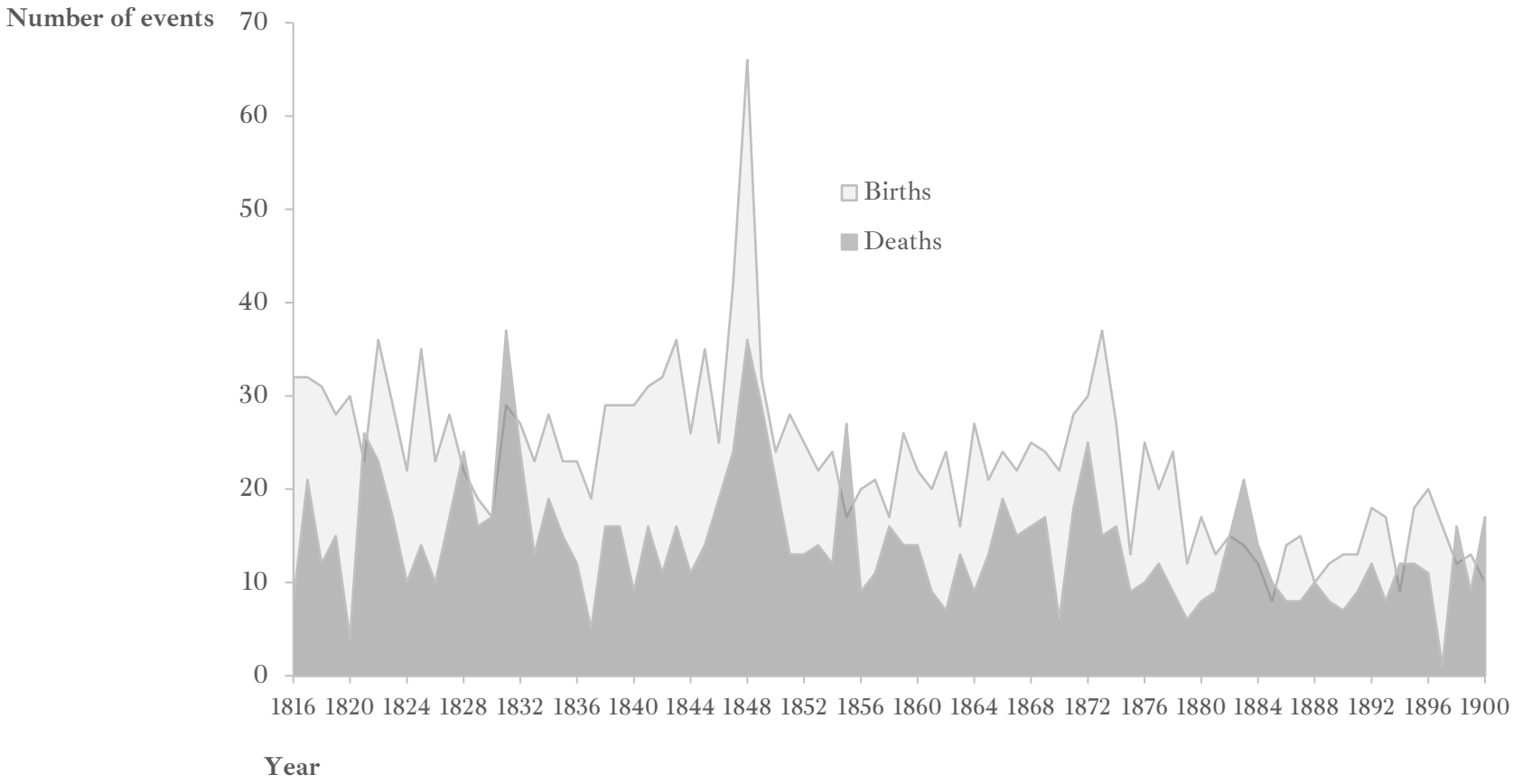
Source: own work based on Archiwum Państwowe w Poznaniu Oddział w Pile, Akta miasta Trzcianka, sygn. 55/15/0/9/300.

Figure 7. Vital statistics for the Jewish community in Trzcianka, 1816–1900



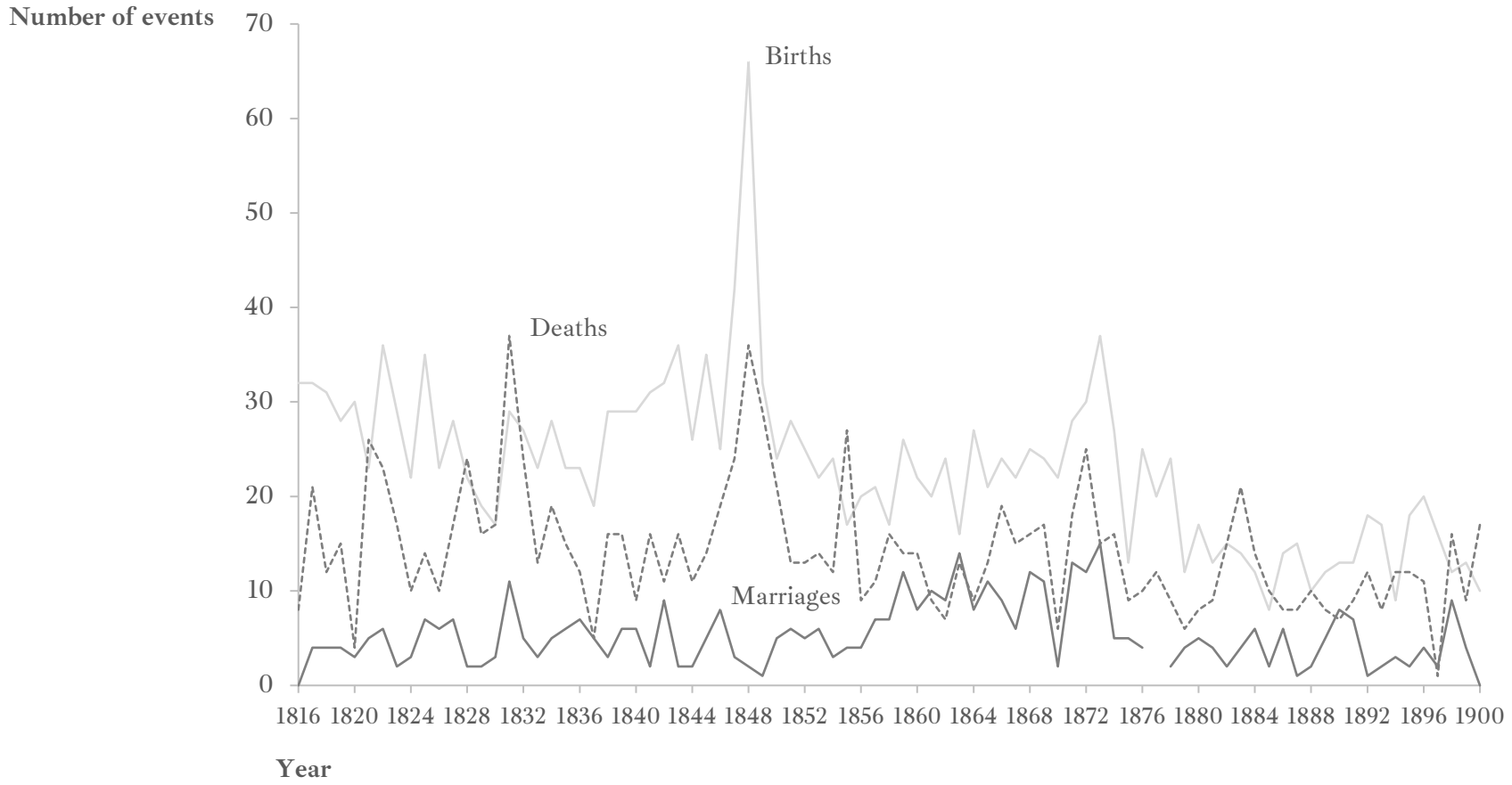
Source: own work based on Moses Löb Bamberger, *Geschichte der Juden in Schönlanke* (Berlin: Verlag von Louis Lamm, 1912), 34.

Figure 8. Vital statistics for the Jewish community in Trzcianka, 1816–1900



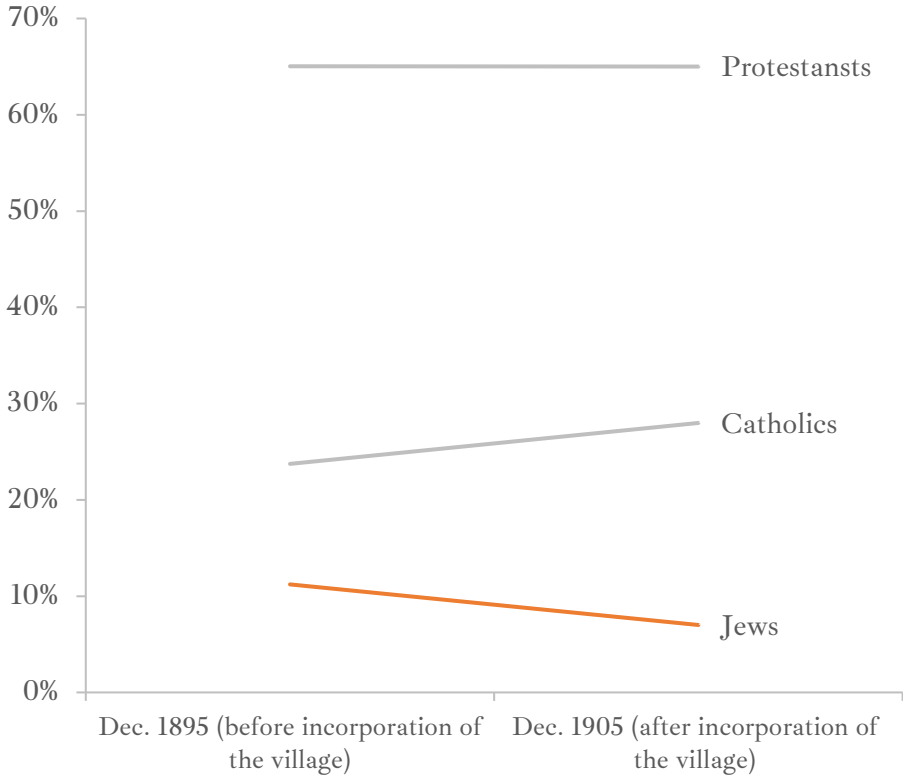
Source: own work based on Moses Löb Bamberger, *Geschichte der Juden in Schönlanke* (Berlin: Verlag von Louis Lamm, 1912), 34.

Figure 9. Vital statistics for the Jewish community in Trzcianka, 1816–1900



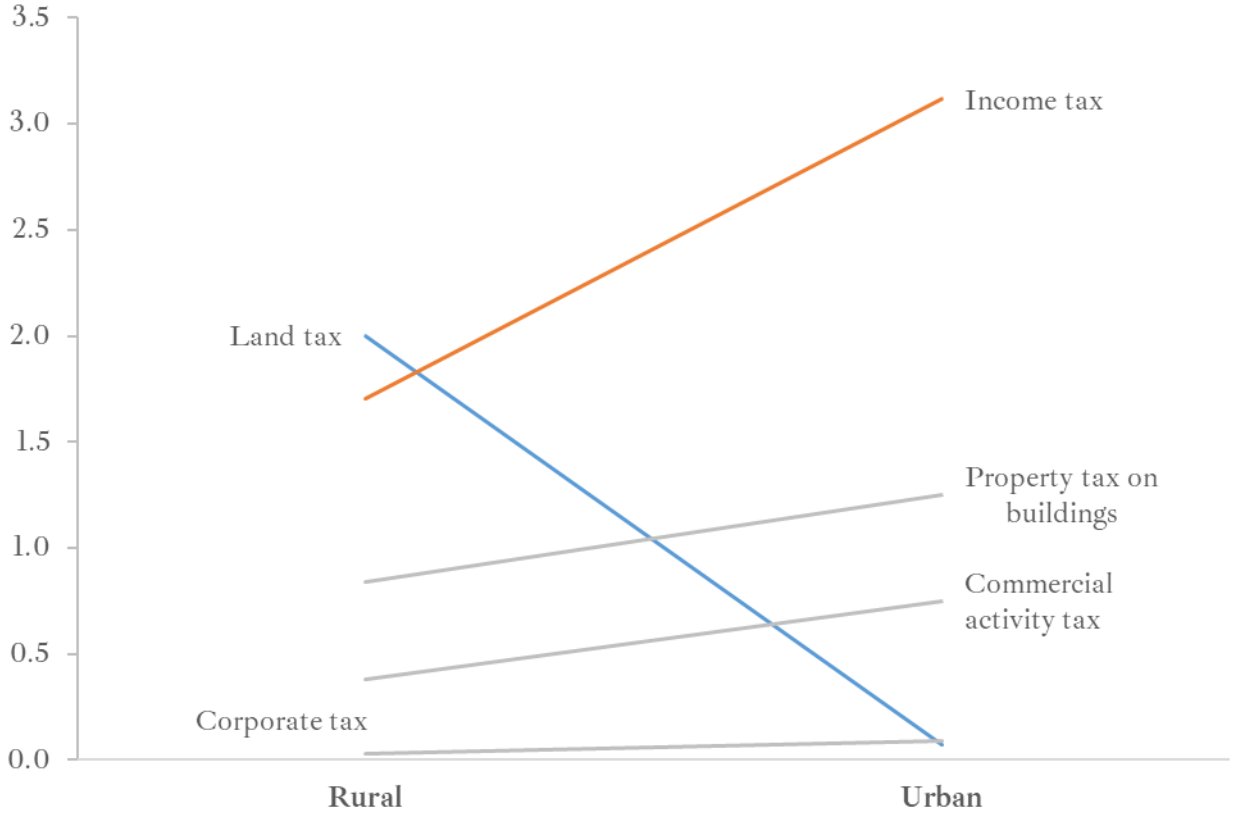
Source: own work based on Moses Löb Bamberger, *Geschichte der Juden in Schönlanke* (Berlin: Verlag von Louis Lamm, 1912), 34.

Figure 10. Population of the town of Trzcianka in 1895 and 1905, by religion



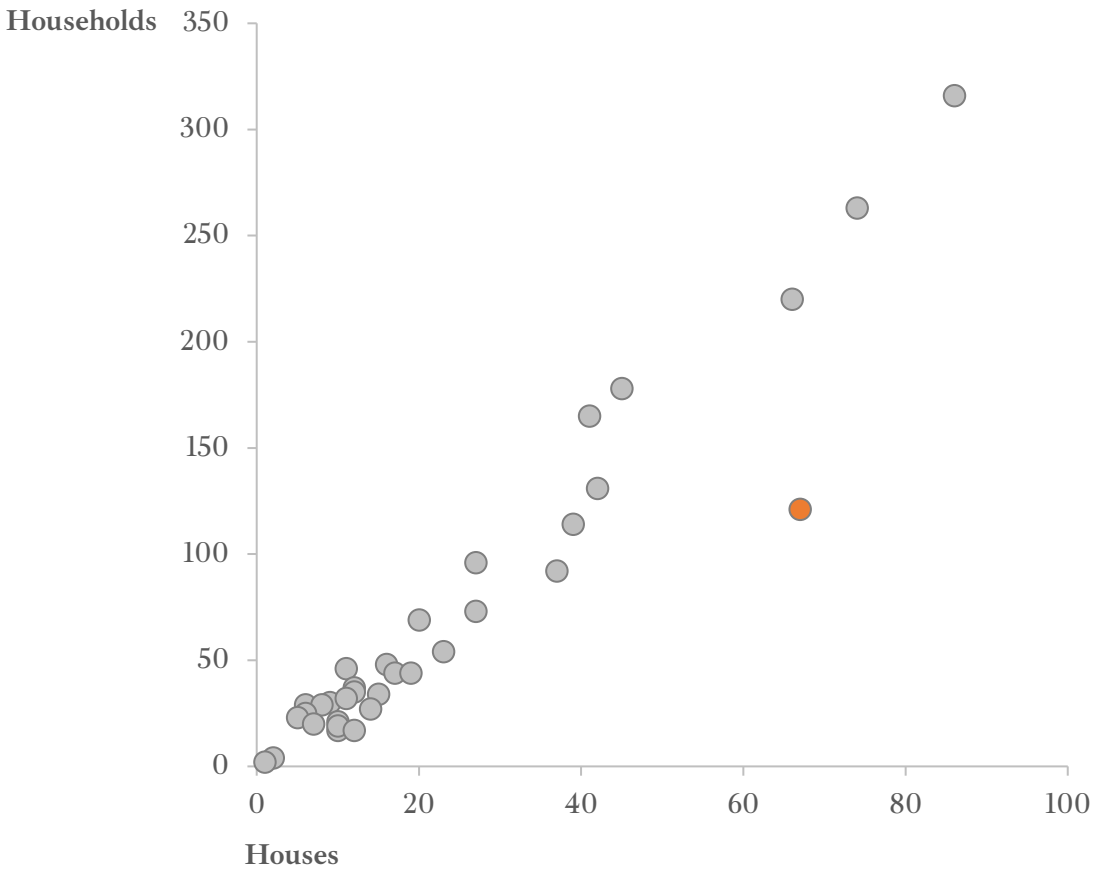
Source: own work based on *Gemeindelexikon für das Königreich Preussen. Auf Grund der Materialien der Volkszählung vom 2. Dezember 1895 und anderer amtlicher Quellen*, Band 5: *Provinz Posen* (Berlin: Verlag des Königlichen statistischen Bureau, 1898), 172–173; *Gemeindelexikon für das Königreich Preußen. Auf Grund der Materialien der Volkszählung vom 1. Dezember 1905 und anderer amtlicher Quellen*, Heft V: *Provinz Posen* (Berlin: Verlag des Königlich Preussischen Landesamte, 1908), 21–23.

Figure 11. Tax amount per capita (in marks) for 1905, in urban and rural Trzcianka



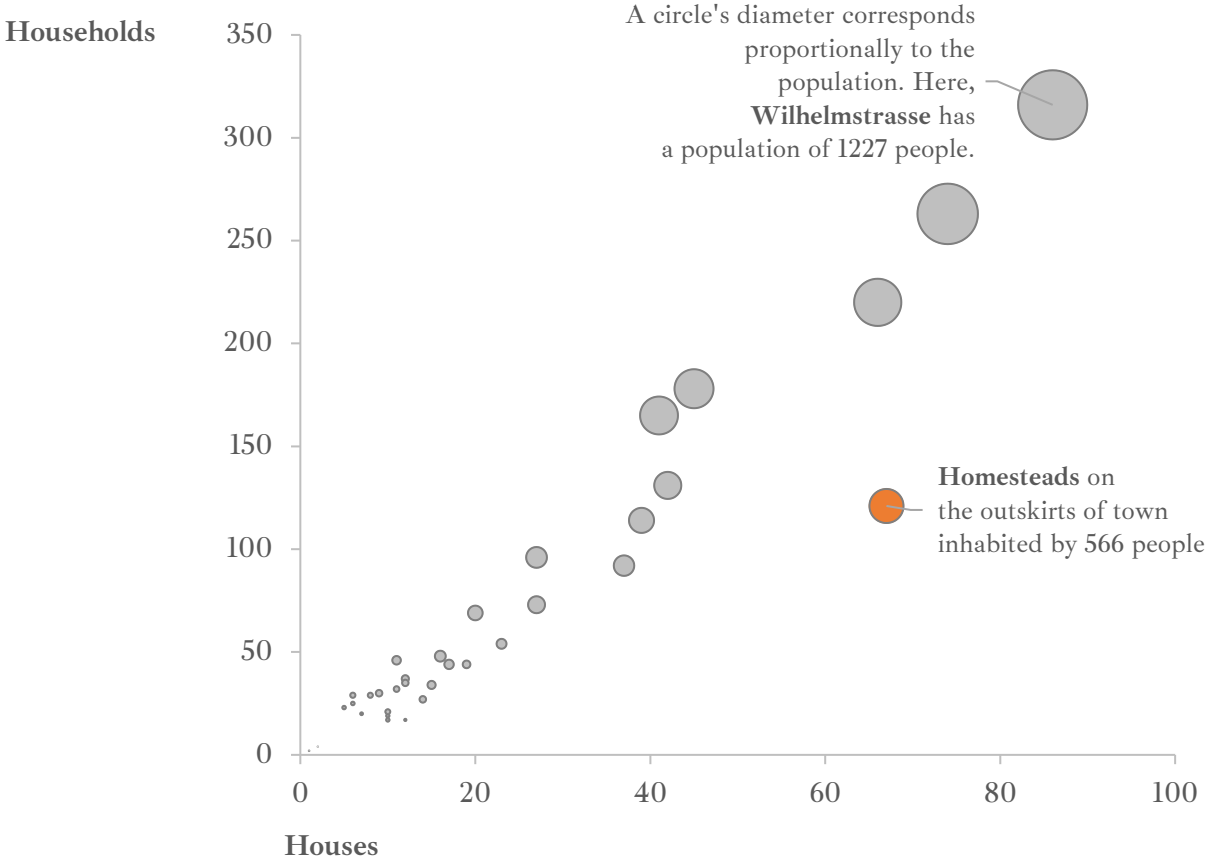
Source: own work based on Geheimes Staatsarchiv Preussischer Kulturbesitz, I HA. Rep. 77 Ministerium des Innern, Tit. 2686 Nr. 5.

Figure 12. Number of houses and households on each street in the town of Trzcianka in 1927



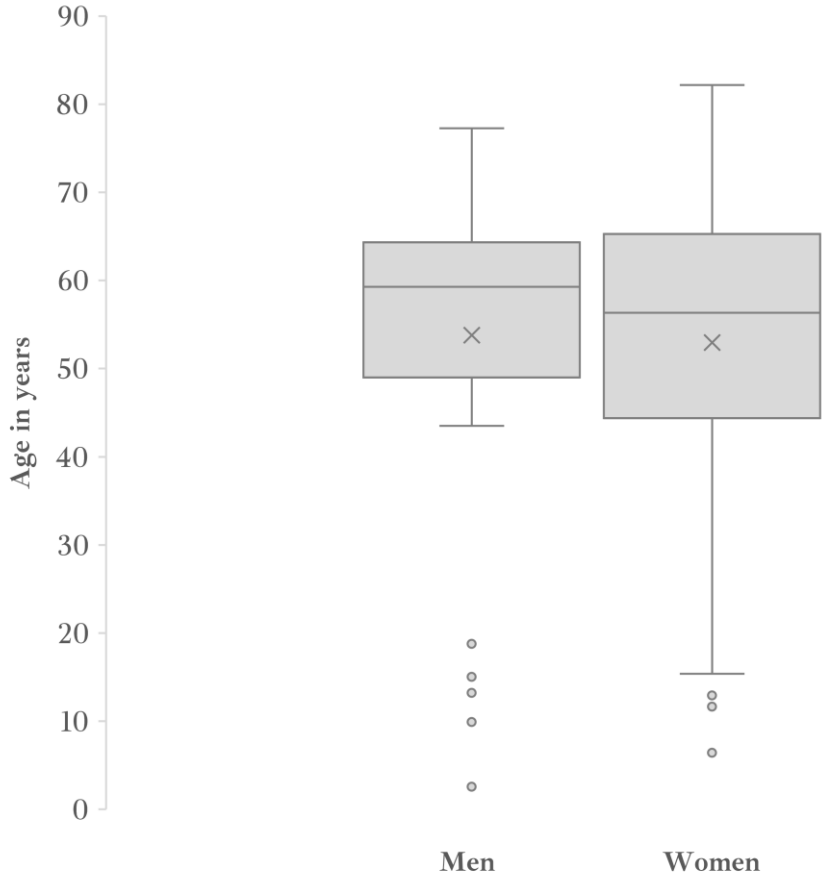
Source: own work based on Archiwum Państwowe w Poznaniu Oddział w Pile, Akta miasta Trzcianka, sygn. 55/15/0/3/518.

Figure 13. Number of houses and households weighted by population on individual streets in the town of Trzcianka in 1927



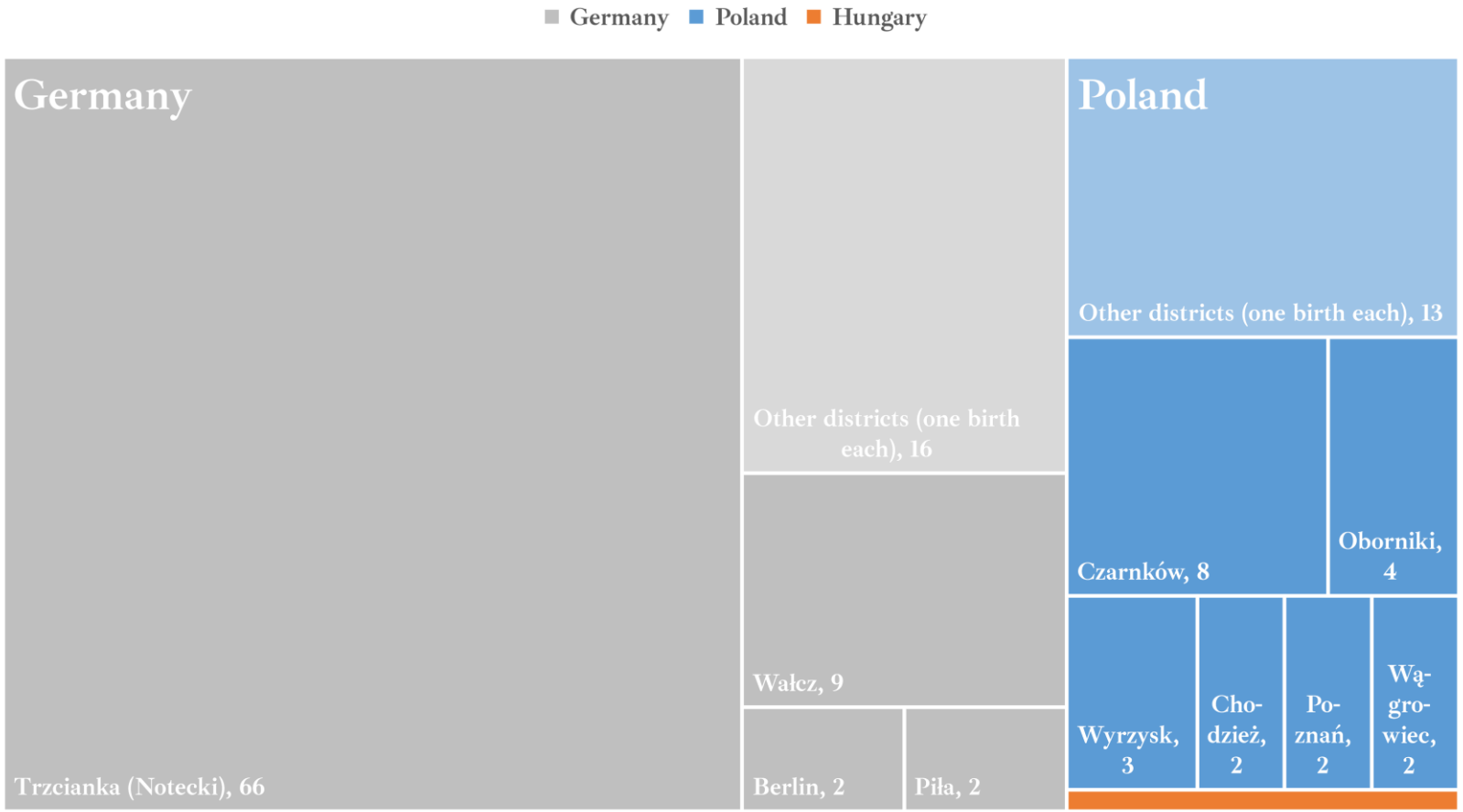
Source: own work based on Archiwum Państwowe w Poznaniu Oddział w Pile, Akta miasta Trzcianka, sygn. 55/15/0/3/518.

Figure 14. Age of population recognized as Jewish in Trzcianka on May 17, 1939, by gender



Source: own work based on *Mapping the Lives. A Central Memorial for the Persecuted in Europe 1933–1945*, accessed 12.01.2023, hhttps://www.mappingthelives.org.

Figure 15. Number of Jewish inhabitants in Trzcianka on May 17, 1939 by place of birth (countries, districts; borders in 1937)



Source: own work based on *Mapping the Lives. A Central Memorial for the Persecuted in Europe 1933–1945*, accessed 12.01.2023, hhttps://www.mappingthelives.org.