Data Journalism – Impact of Statistical Methods

Claus Weihs, Marcel Pauly and Patrick Stotz

Abstract Data journalism strongly depends on adequate data preparation and analysis. In this paper, we discuss the impact of statistical methods on data journalistic analysis. To this end, we re-analyze two data journalistic publications of SPIEGEL ONLINE with more advanced statistical methods and discuss pro and contra.

Claus Weihs
TU Dortmund University, D-44221 Dortmund,
✉ claus.weihs@tu-dortmund.de

Marcel Pauly · Patrick Stotz
SPIEGEL ONLINE
✉ marcel.pauly@spiegel.de
✉ patrick.stotz@spiegel.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 6, No. 2, 2020

DOI: 10.5445/KSP/1000098012/15
ISSN 2363-9881

1 Introduction

Data journalism is a strongly emerging journalistic discipline. Besides the acquisition and preparation of data, data journalistic publications are strongly based on adequate data analysis. Most often, visualization methods are used for such analyses. In this paper, we will discuss the impact of statistical methods on the quality of outcomes.

There appear to be only few scientific studies about the quality of statistical evaluations or data visualizations in journalism (see, e.g. Young et al. (2018) and the references therein). From a practical viewpoint, data journalists repeatedly observe that many relevant analytical methods are terra incognita, and although the quality of data journalistic work appears to have been clearly improved in the last years, even today we quite often find publications with methodological deficiencies. For example, in data journalistic textbooks and tutorials one still finds basic advices about the usage of medians, means etc. and that one should not confuse correlation with causality (see, e.g., Cairo (2016)).

Having in mind that magazine readers very often represent the general public with totally different skills in reading comprehension as well as mathematical/statistical knowledge, data journalistic analyses are supposed to be fairly simple and easily comprehensible. For example, data journalists sometimes assume that even the interpretation of scatterplot diagrams can lead to problems of comprehension. While the benevolent data journalist does want to exclude misinterpretations, he or she also wants to show the most relevant of what can be learned from the available data. Therefore, advanced methods sometimes appear to be necessary. This creates the need for simple explanations of advanced methods and, especially, their results. This is one of the main motivations for this paper.

To this end, we will study two data journalistic projects: “Age structure of parliamentarians and population” and “social indicators in party strongholds”, both first published on SPIEGEL ONLINE in German. For this paper, the SPIEGEL ONLINE results are translated to English. For both projects, we first give a brief project description with a presentation of the results of the published SPIEGEL ONLINE article and then the statistician’s view on the project in a re-analysis. For the re-analyses the software R is used (R Core Team, 2018).

2 Age Structure of Parliamentarians and Population

In 2017, SPIEGEL ONLINE analyzed the question "How much represent parliamentarians the whole population demographically?" (Segger et al., 2017).

2.1 SPIEGEL ONLINE Analysis

SPIEGEL ONLINE used data from 23 European countries including the population distribution according to age groups and sex (from Eurostat) and sex and age of all parliamentarians (from diverse sources). As visualization, age pyramids are used (Figure 1 for Germany). For the comparison of the female share of parliamentarians, even more countries are included (Figure 2).
Considering the German age distributions (Figure 1), obviously, in nearly every age group, men are over- and women are under-represented in parliament. Moreover, young and older people are under-represented. There are only two countries where men are not over-represented, namely in Rwanda and Bolivia (Figure 2).

2.2 Re-Analysis

Statisticians' critics on SPIEGEL ONLINE's data analysis is based on the following arguments:

- Frequencies in age pyramids relate also to non-eligible people. Instead, conditional distributions should be used.
- Distances of age distributions are only qualitatively compared. Instead, distance measures for distributions should be used.
- Ranking of the countries is only based on one indicator (difference of female shares in parliaments and population). Instead, ranking should be based on distribution distance measures.

For the re-analysis, SPIEGEL ONLINE kindly provided their data. Using conditional distributions restricted to the age groups from 20 years on, i.e. excluding all age groups with non-eligible people, obviously decreases the

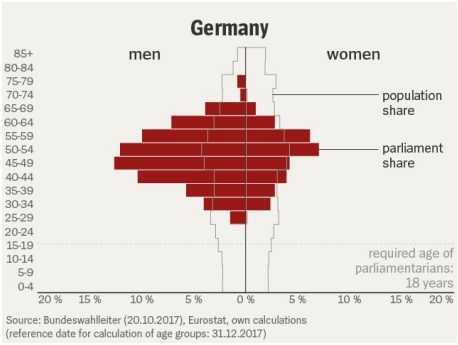

Figure 1: Stacked age pyramid for Germany; parliament share vs. population share, men vs. women.

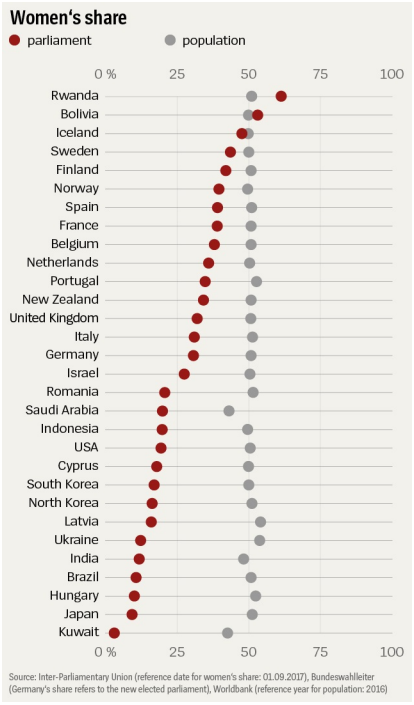

Figure 2: Women's share in parliaments and population for different countries.

distance of the distributions of parliamentarians and population (Figure 3). From the age pyramids, one can see that the great majority of parliamentarians is between 20 and 69 years old. Therefore, in the following representations we only include age groups between 20 and 69, also for the population. Note that in this way, percentages of men and women together do not sum up to 100 %, since there are parliamentarians and (especially) population parts who are older than 69 years.

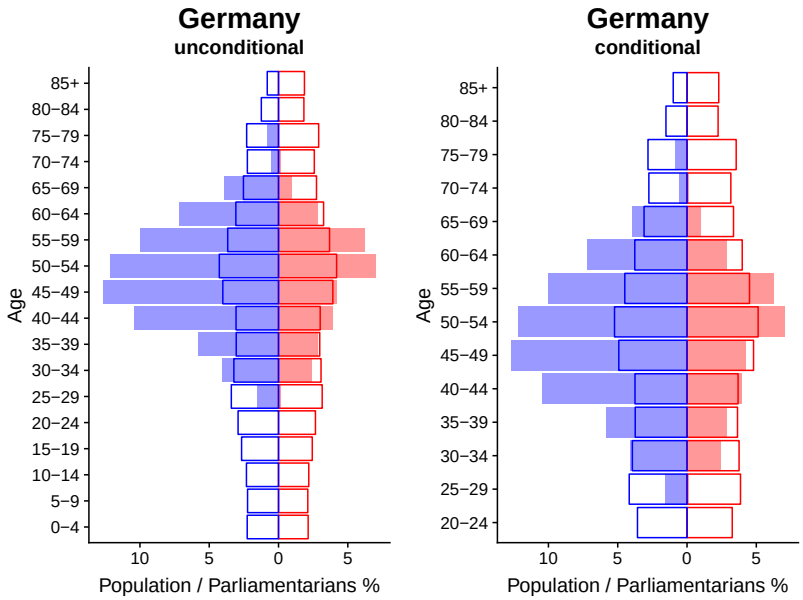

Figure 3: Stacked age pyramids for Germany: unconditional (left) vs. conditional (right); blue = male, red = female, open = population, filled = parliamentarians.

Now, we change the representation from age pyramids to densities and distribution functions, as usual in statistical analysis. (Empirical) densities are age pyramids rotated by 90 °, individually for men and women, with relative frequencies / percentages of age groups (Figure 4, bottom). As an alternative, (empirical) distribution functions are used, for which shares are stepwise summed up from the youngest to the oldest age group (Figure 4, top). Both representations together show, e.g., that women between 50 and 59 years are over-represented in parliaments, though women are overall under-represented.

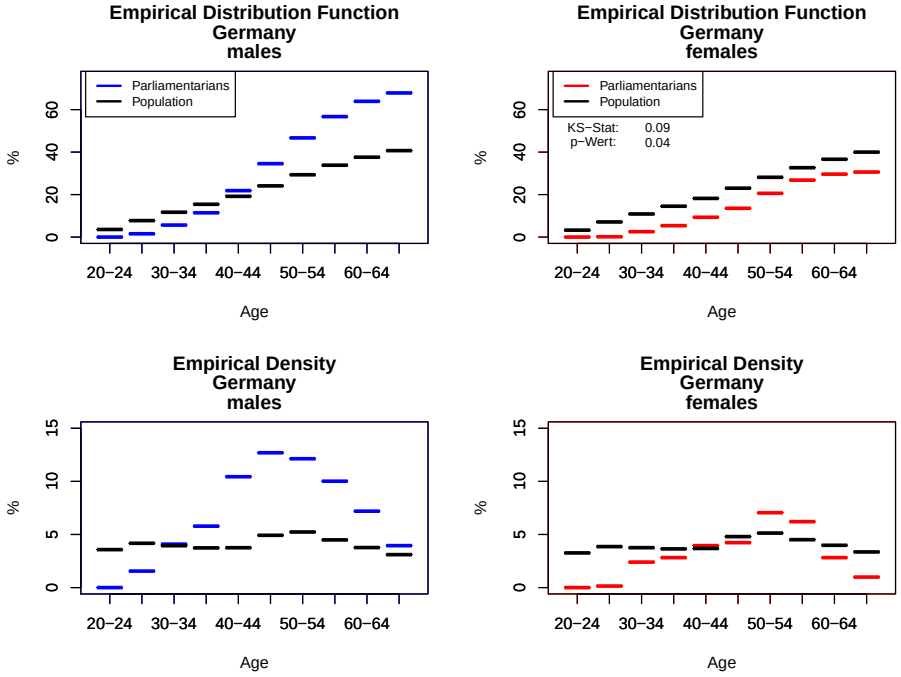

Figure 4: Empirical distribution functions (top) and empirical densities (below) for Germany; blue = male parliamentarians, red = female parliamentarians, black = correSp. population.

SPIEGEL ONLINE only compared women's shares in parliaments and populations (Figure 2). For an objective assessment of the difference between two distributions, distance measures between the whole distributions should be used based on distribution functions or densities. The idea is to adapt distance measures from Geometry for two n-dimensional points $(x_1\, x_2 \ldots x_n)$ and $(y_1\, y_2 \ldots y_n)$ (Table 1, top). We define analogue distance measures for distributions, one based on distribution functions and two based on densities (Table 1, bottom). We apply these measures to the distribution function / density values of the n = 10 age groups [20,24], [25,29], ..., [60,64], [65,69], i.e. to the values $F_1,\ F_2, \ldots, F_{10}$ and $G_1,\ G_2, \ldots, G_{10}$ of two distribution functions or $p_1,\ p_2, \ldots, p_{10}$ and $q_1,\ q_2, \ldots, q_{10}$ of two densities (relative frequencies). Note the weighting in the χ^2-distance with R and S being the total numbers of observations, the p_i and q_i are based on, respectively (Press et al., 1992).

Table 1: Distances in Geometry (top) and between Distributions and Densities (bottom).

Maximum distance	$\max_{i=1,\ldots,n} \lvert x_i - y_i \rvert$
Sum distance	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Euclidian distance	$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
Kolmogorov-Smirnov (KS-)distance	$\max_{i=1,\ldots,n} \lvert F_i - G_i \rvert$
Total variation	$0.5 \sum_{i=1}^{n} \lvert p_i - q_i \rvert$
χ^2-distance	$\sqrt{\sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i/S + q_i/R}}$

With the above distance measures, we can compare the distribution of the parliamentarians (male or female) to the distribution of the relevant population (for Germany, the relevant distributions are given in Figure 4). Such distances can be used to determine a ranking of different countries. Comparing Germany with, e.g., Sweden and Romania we get the ranking in Table 2 based on the Kolmogorov-Smirnov (KS-) distance. Note that for females, a different ranking results from the total variation and χ^2-distances (rankings in parentheses).

Table 2: Rankings.

Country	Sex	KS-Dist. (rank)	total var. (rank)	χ^2 dist. (rank)	difference in female's share
Sweden	male	0.13 (1)	0.13 (1)	32 (1)	
Germany	male	0.27 (2)	0.20 (2)	144 (2)	
Romania	male	0.35 (3)	0.24 (3)	173 (3)	
Sweden	female	**0.07** (1)	0.10 (2)	23 (2)	4 (1)
Germany	female	0.09 (2)	**0.09** (1)	24 (3)	20 (2)
Romania	female	0.21 (3)	0.12 (3)	**11** (1)	31 (3)

For females the smallest KS-distance is realized by Sweden and the smallest χ^2-distance by Romania (in bold). Germany's KS-value is also not very big, but the distribution systematically lies below the distribution of the total population (see Figure 4). For comparison, see the distribution for females in Sweden in Figure 5. Note that differences between countries appear to be smallest for total

variation, but much bigger using SPIEGEL ONLINE's difference of female's share in population and parliament (Table 2, last column).

Obviously, the choice of distance measures can determine the results. Note, however, that this choice also determines the way to assess the distance. For example, KS only looks for the maximum difference, for total variation all distances are weighted equally, and for the χ^2-distance squared distances are weighted differently.

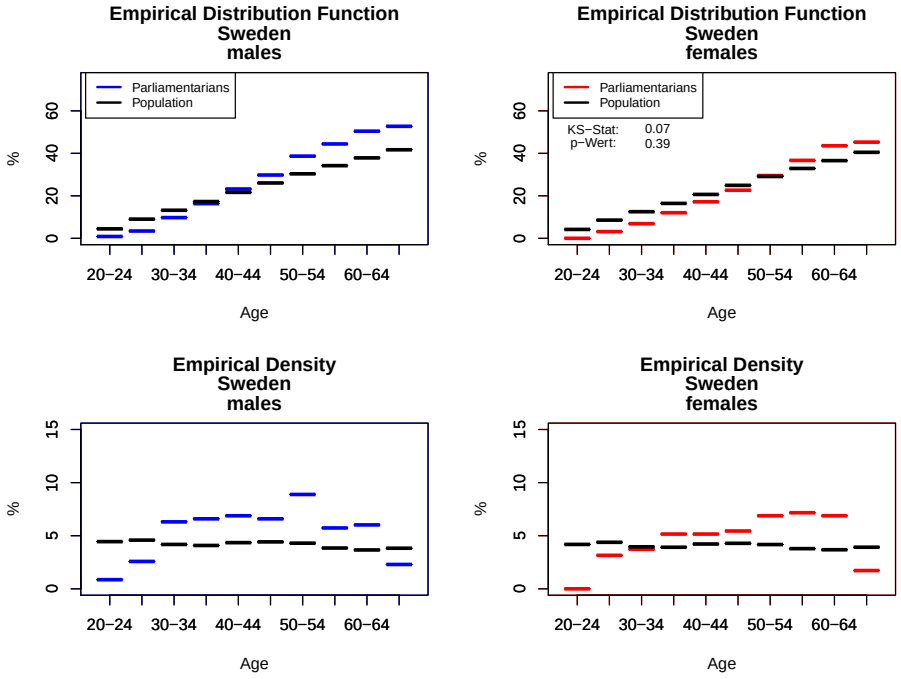

Figure 5: Empirical distribution functions (top) and empirical densities (bottom) for Sweden.

To summarize, we used conditional distributions for the comparison of the parliamentarians' distribution with the relevant part of the population. We used densities and distribution functions as statistical alternatives to age pyramids. Distances between distributions are used as objectifications of distance. Rankings

are based on such distances. The choice of such a distance measure appears to be crucial, but not straight forward.

A possible extension are statistical tests trying to assess whether the difference between the distribution of parliamentarians and the conditional distribution of the population is *significant* with regard to a distance measure. Since it is well known that such tests are not easy to understand by non-statisticians, we abstain to propose them here.

3 Social Indicators in Party Strongholds

Directly after the parliamentary election in 2017, SPIEGEL ONLINE analyzed questions like "What characterizes the voters of the individual parties?" and "How has the voter structure changed compared to 2013?".

3.1 SPIEGEL ONLINE Analysis

Data used were the socio-demographic factors income, unemployment, foreigner share, and population density in the election districts. Naturally, one aim was publication as soon as possible after election and a visually attractive presentation - not only densely packed text with many numbers. Indeed, the article was ready in the night after the election, published by SPIEGEL ONLINE the next afternoon. It was a brief text with 8 diagrams. Density curves were used as the central visual element (Holscher et al., 2017).

From the densities for the strongholds of different parties (best 75 results in the 299 election districts), one can easily derive that the union parties CDU/CSU have their strongholds quite dominantly in election districts with low unemployment rates (Figure 6) and that in 2013 the (right wing) AfD had its strongholds in election districts with average income (and foreigner share), but in 2017 best results were achieved in districts with low income (and low foreigner share, see Figure 7).

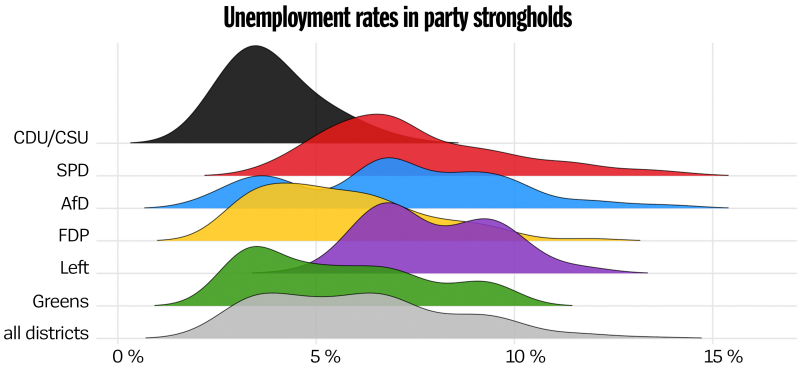

Figure 6: Unemployment rates in party strongholds in 2017.

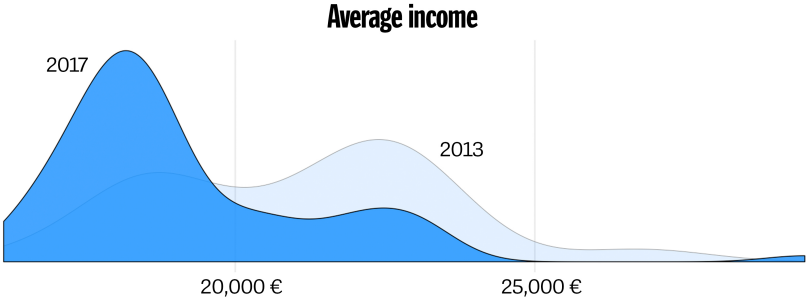

Figure 7: Average income in AfD strongholds 2013 and 2017.

3.2 Re-Analysis

Statisticians' critics on SPIEGEL ONLINE's analysis of the data is based on the following arguments:

– Influences are only considered for single factors. Instead, multiple influences should be considered simultaneously. i.e. classification / regression methods should be tried.

– Predictive quality is not quantified. Therefore, error measures should be introduced.

For the re-analysis we used the same data as SPIEGEL ONLINE (Bundeswahlleiter, 2017a,b). In the re-analysis we restrict ourselves to the prediction of AfD-strongholds in 2017. SPIEGEL ONLINE used the structural features of the election districts only individually (Figure 7). The idea is to predict the AfD-strongholds from the same structural data simultaneously, i.e. from the four features unemployment rate, average income, foreigner share, and population density together. First, we identified the 75 AfD-strongholds having an AfD-percentage higher than 14.2 % and show, as SPIEGEL ONLINE before, that the foreigners' share is relatively low in these districts (Figure 8). Moreover, we realized that most of the high percentages appeared in Eastern Germany (in red in Figure 9).

Therefore, as the classification problem we identified: Separate the two Classes "AfD-Stronghold" (AfD > 14.2 %, class 1) from "not AfD-Stronghold" (class 0), i.e. determine the optimal prediction for the 'true' class from the structural data.

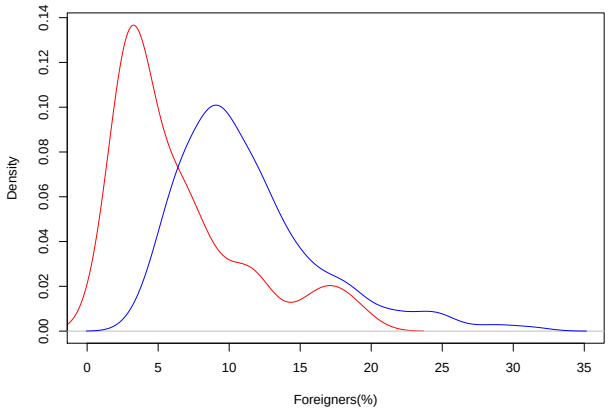

Figure 8: Densities of foreigner share. Red: AfD-stronghold, blue: not AfD-stronghold.

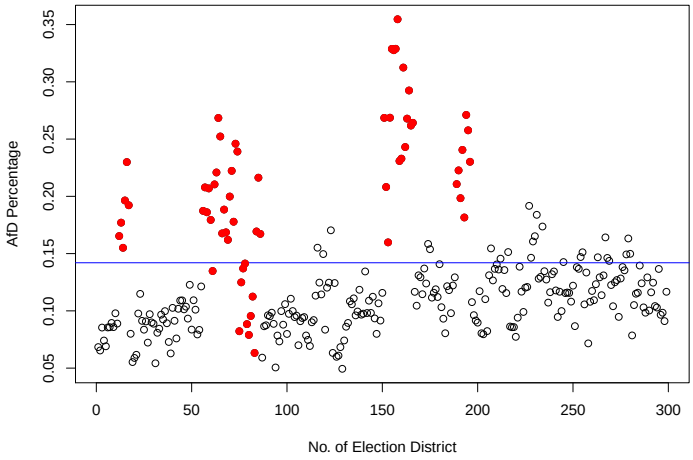

Figure 9: AfD-percentage in election districts. Red: East Germany, above blue line: 75 districts.

There is a plethora of different classification methods, most of them with the disadvantage of bad interpretability of classification rules. We use a method whose rules are directly interpretable, the so-called decision trees.

Classification problems are intrinsically prediction problems. Prediction corresponds here to the unknown class of known structural data, and not to future time points! [1] Error rates and confusion matrices are used to assess the quality of a classification rule.
The *error rate e* is the relative prediction error when applying the rule and the *confusion matrix* compares the true and the predicted classes in more detail.

The re-analysis aims at optimal prediction whether an election district is an AfD-stronghold (more than $14.2\,\%$ AfD-voters, class 1) or not (class 0) on the basis of the structural features unemployment rate, average income, foreigner share, and population density. Since we have seen that many AfD-strongholds are in eastern Germany, we also include an East-Indicator (east = 0, if the election district lies in Western Germany; east = 1 else, incl. Berlin).

[1] In this paper, we predict classes by means of the rule learned by all observations (resubstitution error).

The resulting decision tree in Figure 10 has a prediction error rate of $e = 8\,\%$. Note that in the 'leafs' of the tree the predicted class is given followed by "no. of errors / no. of included districts". The confusion matrix (Table 3) shows that

- not-AfD-strongholds are predicted correctly in more than $98\,\%$ $(220/224 > 0.98)$, but
- AfD-strongholds only in approximately $73\,\%$ $(55/75 = 0.73333)$.

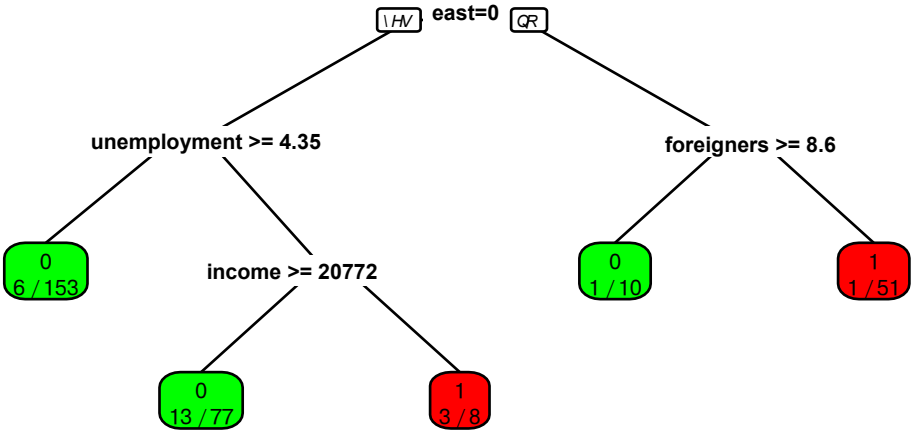

Figure 10: Prediction of AfD-strongholds Decision tree. Green/0: not AfD-stronghold. Red/1: AfD-stronghold.

Table 3: Confusion Matrix.

		Prediction	
		0	1
Class	0	220	4
	1	20	55

The visualization of the decision tree in Figure 10 can be transformed into simple colloquial rules as follows:

Rule 1: In an East-district, an AfD-stronghold is very probable (in $98\,\% = 50/51$) if the foreigner share is lower than $8.6\,\%$. (Cp. SPIEGEL ONLINE)!

Rule 2: In a West-district, a not-AfD-stronghold is very probable (in $96.1\,\%$ = $147/153$) if the unemployment rate is greater or equal $4.35\,\%$.

Rule 3: In a West-district, a not-AfD-stronghold is probable ($83.1\,\%$ = $64/77$) if the unemployment rate is lower than $4.35\,\%$ and the average income greater or equal $20772\,€$.

Rule 4: In a West-district, an AfD-stronghold is only then relatively probable ($62.5\,\%$ = $5/8$) if the unemployment rate is lower than $4.35\,\%$ and the average income lower than $20772\,€$.

Such rules could (together with the error rate and the interpretation of the confusion matrix) be discussed in a press report, with or without showing the tree. Note that the feature 'population density' is not included in the decision tree since the other features appear to be much more important for class separation.

One might argue that a prediction error rate of $8\,\%$ is too high and try to build an extended tree with more features and a lower error rate. Indeed, in the structural data there were more features which could be utilized for prediction, and the prediction error rate can be decreased to $e = 6\,\%$, if, additionally, the following features are used: "No. of motor vehicles per 1000 inhabitants" (as a prosperity indicator) and "Share of inhabitants 60 years or older". However, in this way, the tree (Figure 11) gets much more complex at the "western side" and much less interpretable. Therefore, the original tree might be preferred. Note that, on the one hand, it is obvious that an extended tree might model the data better, but that, on the other hand, it might also model noise in the observations so that the error estimate is more uncertain. This is called the bias-variance trade-off. For a discussion of this trade-off see, e.g., Schiffner et al. (2012).

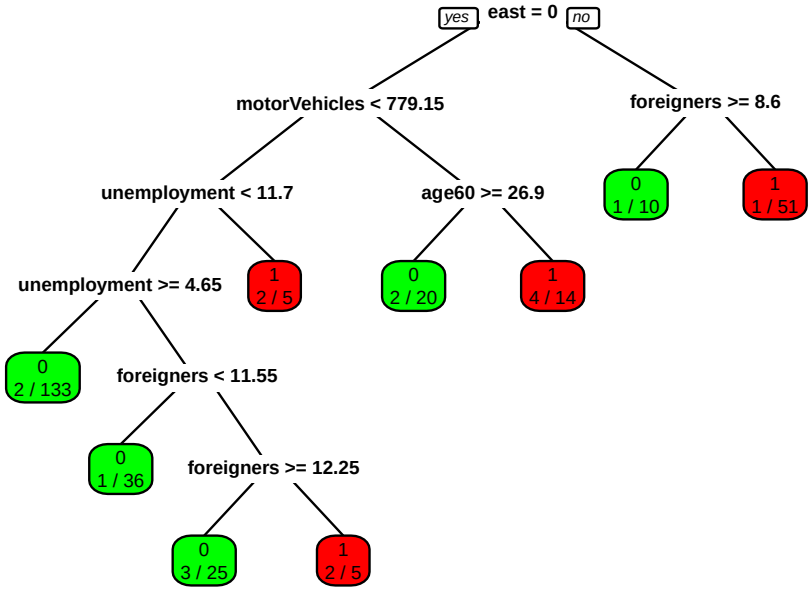

Figure 11: Extended prediction of AfD-strongholds. Green/0: Not AfD-stronghold. Red/1: AfD-stronghold.

To summarize, as methods for the re-analysis of SPIEGEL ONLINE's results we utilized classification based on the joint use of all available indicators, decision trees providing attractive visualization and clear and easy rules, as well as error rates and the confusion matrix as measures for classification quality.

4 Conclusion

In one study, we proposed conditional distributions and rankings based on distances between distributions. In the other study, we proposed classification with decision trees based on the joint use of all available indicators as well as error rates and the confusion matrix as measures for classification quality.

Based on this, we discussed the question: Are these statistical methods / results understandable for the 'representative reader'? Some comments of the involved journalists were:

- Distance measures could be helpful in an investigation to build rankings and identify newsworthy cases. In order to use them in journalistic articles, though, their meaning has to be clearly and briefly explained in words.
- On the one hand, the simultaneous usage of several features allows for more detailed conclusions and better interpretation of election results. On the other hand, the quick communication of graphics might be more complicated and the contents might be misunderstood by readers.

Acknowledgements We would like to thank Dr. Daniel Horn for his critical comments which definitely improved the paper.

References

Bundeswahlleiter (2017a) Bundestagswahl 2017: Endgültige Ergebnisse. URL: https://www.bundeswahlleiter.de/bundestagswahlen/2017/ergebnisse.html.

Bundeswahlleiter (2017b) Bundestagswahl 2017: Strukturdaten für die Wahlkreise. URL: https://www.bundeswahlleiter.de/bundestagswahlen/2017/strukturdaten.html.

Cairo A (2016) The Truthful Art: Data, Charts, and Maps for Communication (Voices That Matter). New Riders, San Francisco. ISBN: 978-0-133440-49-2.

Holscher M, Segger M, Stotz P (2017) Kaum Ausländer in AfD-Hochburgen - Union besonders auf dem Land beliebt. URL: https://www.spiegel.de/politik/deutschland/bundestagswahl-2017-kaum-auslaender-in-afd-hochburgen-a-1169727.html.

Press W, Teukolsky S, Vetterling W, Flannery B (1992) Numerical Recipes in C. Cambridge University Press. URL: https://www2.units.it/ipl/students_area/imm2/files/Numerical_Recipes.pdf.

R Core Team (2018) R: A Language and Environment for Statistical Computing. Vienna, Austria. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Schiffner J, Bischl B, Weihs C (2012) Bias-Variance Analysis of Local Classification Methods. In: Challenges at the Interface of Data Analysis, Computer Science, and Optimization, pp. 49–57. Studies in Classification, Data Analysis, and Knowledge

Organization. Springer, Berlin, Heidelberg, Gaul W, Geyer-Schulz A, Schmidt-Thieme L, J. K (eds.). DOI: 10.1007/978-3-642-24466-7_6.

Segger M, Pauly M, Stotz P (2017) Wo Bolivien und Ruanda den Deutschen Bundestag abhängen. URL: https://www.spiegel.de/politik/deutschland/bundestag-frauenanteil-nur-mittelmass-im-weltweiten-vergleich-a-1174318.html.

Young ML, Hermida A, Fulda J (2018) What Makes for Great Data Journalism? A content analysis of data journalism awards finalists 20122015. Journalism Practice 12(1):115–135. DOI: 10.1080/17512786.2016.1270171.