Geographical Voting Patterns in Berlin: a Brief Data Science Case-Study.

Brendan Osberg

(Dated: June 23, 2022)

Abstract

Voting data sets are a convenient platform for data science: they are quantifiable, controlled, and they exhibit the right balance of noise and signal. In Germany, in particular, the compulsory 'Wohnmeldungs' allow for very detailed information in each riding. Moreover, in Berlin (as well as at the national level), the historical East/West divide provides a prominent feature for binary classification. The results shown in this report represent the results of playing with this data a bit, primarily as a pedagogical tool for data science methodology.

That an ideological divide remains between districts from the East and West has long been known qualitatively, however I want to quantify it more precisely here and demonstrate that integration has actually succeeded in some demographics. I start by performing cluster analysis on the N =1,779 districts of the city according to their vote-distribution among the major parties; *three* clusters are observed.

Two of the clusters, as expected, were overwhelmingly dominated by points east and west of the former wall respectively. A third cluster, however, was distributed almost exactly according to the null distribution of overall districts –indicating that the barrier no longer has ideological significance to voters there. Where are these districts? More to the point: where exactly have efforts at integration been particularly successful?

It turns out that it is in these regions where the Green party derives much of its support. What is it that the Green party are doing particularly well, compared to other parties in appealing to moderate (or 'integrated') voters from both sides of the former wall?

I. METHODOLOGY:

- I took publicly available data from the 1779 voting districts in Berlin from the 2016 election; Votes carried out by mail showed the same trend but lower sample size, thus we focus on the "Urnenwahlbezirk"s.

- All parties that received less than 5% of the total vote were eliminated from the analysis. This preserved the D =6 'major' parties in the following order: SPD, CDU, Green, Linke, AfD, and FDP.

- Each district was then assigned a 'position' in the D-dimensional vector space of popular support for the major parties, without any regard to its geographical location. The proximity of points in this space can then be considered a measure of ideological consistency (i.e. two districts that are 'close' to one-another in this vector space had a similar voting distribution among the six parties, regardless of their geographical location.)

- We use the subscript $m \in [1, N]$ to index the riding (where N =1779), and the superscript $i \in [1, D]$ to index the political party, and we define a vector v_m^i characterizing the vote shares assigned to each party i in bezirk m. Normalization implies that

$$\sum_i^D v_m^i = 1 \tag{1}$$

for all v_m. Furthermore, we can define the ideological vector (along various party lines) separating riding m and n, as well as the magnitude of the distance separating them, using the Euclidean norm [1]

$$\vec{u}_{mn} = \vec{v_m} - \vec{v_n} \tag{2}$$

$$|u_{mn}| = \sqrt{\sum_i (v_m^i - v_n^i)^2}. \tag{3}$$

[1] One might also consider KL divergence as a measure of distance since the elements aren't independent. Anecdotal observation suggests this is unlikely to make a significant difference.)

It is these distances that we use to define how the various voting bezirks are positioned relative to one-another in this D-dimensional ideological space.

A. What patterns emerge from this data?

The first question we wish to address: is there an ideological division between the East and West Berlin ridings in this ideological space? By taking a scatter plot of the points in the vector space described above, Fig. 1, and the dimensionality-reduced Fig. 2 illustrate rather clearly that there is. This conclusion is not novel; newspapers reported on it extensively, but we can quantify it more precisely.

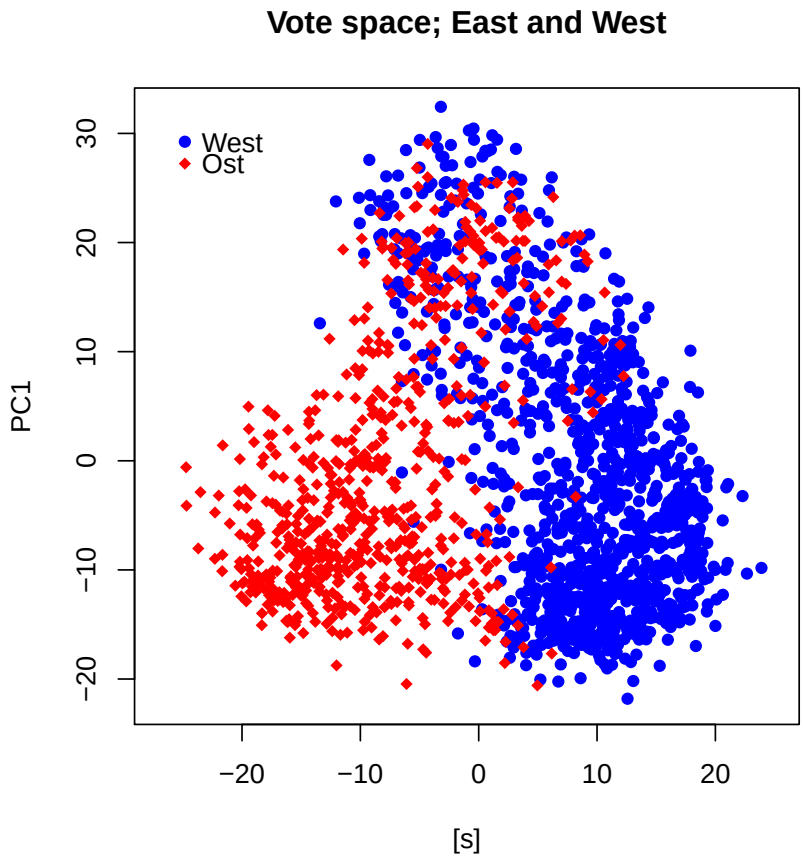

FIG. 1: Popular vote shares by district. Each point is one of the 1,779 districts, and the horizontal axis is defined by the average difference between East and West ridings. Blue points are West, red points are East, with location representing support among the major parties (the vertical axis is the first principle component among remaining axes). Units still correspond to percentage differences.

- From Fig. 1, it is clear that some division between eastern and western districts exist, and yet it is not clear whether data points near the top of the figure represent well-integrated districts, or whether there is some further separation along a dimension orthogonal to the plane.

  For this reason, the data are plotted again in a t-SNE plot in Fig. 2, which reduces the set of full eight-dimensional distances between points into a two-dimensional separation. The same data points from the top of the previous figure are shown well mixed in the bottom-right corner of Fig. 2, indicating that there is, indeed a set of 'well integrated' points.

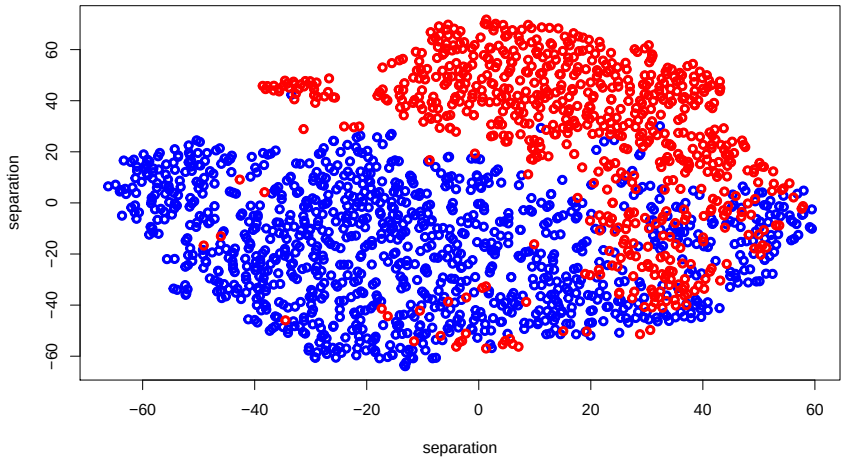

FIG. 2: t-SNE plot of the same data from Fig. 1 color-coded red/blue for east/west. A t-SNE plot collapses the data into a two-dimensional figure where the lateral distance between each point m and n in the 2-D figure approximates $|\vec{u}_{nm}|$ the distance between the same two points in the full D-dimensional space.

B. clustering.

- To further emphasize this point, cluster analysis of these points was then performed using $k-$means clustering for various values of k. When k was set to 2,4,5,6... it was observed that the clusters were not robust: stochastic initial conditions led to random assignment of the data points into each cluster[2]. With $k = 3$, however, clusters were

[2] Here one might be interested in the Jacard index.

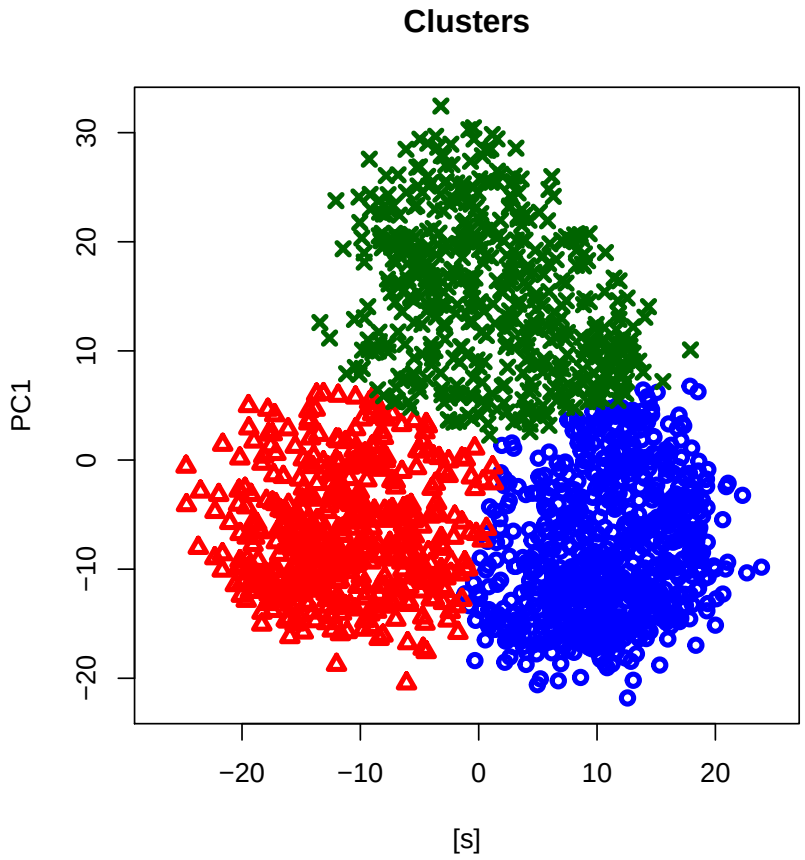

FIG. 3: The same data from Fig. 1 with data points color-coded by k-means clustering. When the 'Integrated' cluster points are omitted, the separation between east and west dominated clusters increases to 25 percentage points.

robust. In this case, we observed an 'Eastern' cluster that was dominated by over 90% ridings east of the wall; a 'Western' cluster was, likewise, dominated by over 90% ridings west of the wall. However, an 'integrated' cluster was also observed, comprising of 65% ridings from the West and 35% ridings from the East, –corresponding very closely to the null expectation of 60.4% (Western) and 39.6%(Eastern) districts overall.

Finally, Fig. 5 shows a bargraph of support for each major party by cluster. The SPD performs approximately equally well everywhere; The Eastern group tends to be polarized between East and West (based on high support for AfD and Die Linke), whereas the Western cluster is where the CDU performed particularly well. The Green party performed by far the best in the regions of the city where the East-West division was no longer significant.

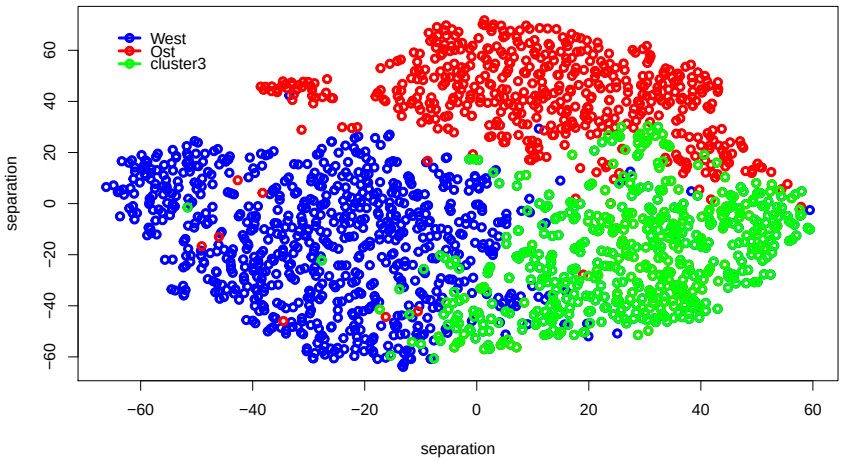

FIG. 4: The same data as in 2, but with cluster 3 (the integrated region) color-coded green Note to self: do further color-coding to show the Eastern districts polarization into hard-left vs. hard-right .

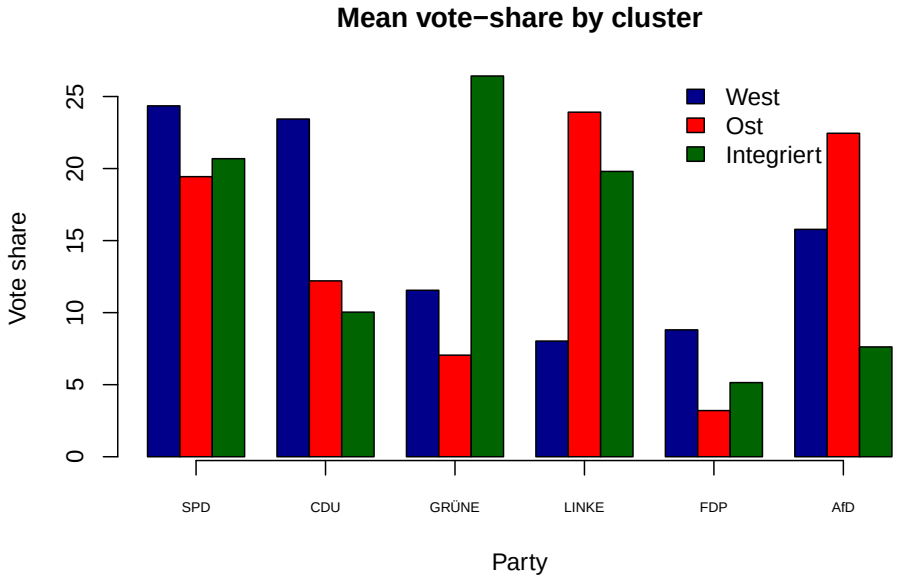

FIG. 5: Mean vote share among the major parties for each cluster.

II. MAPPING SUCCESSFUL INTEGRATION

So if we can see that there is a region where political integration is particularly successful, then where are these regions? Fig. 7 shows the location of the above clusters on a map.

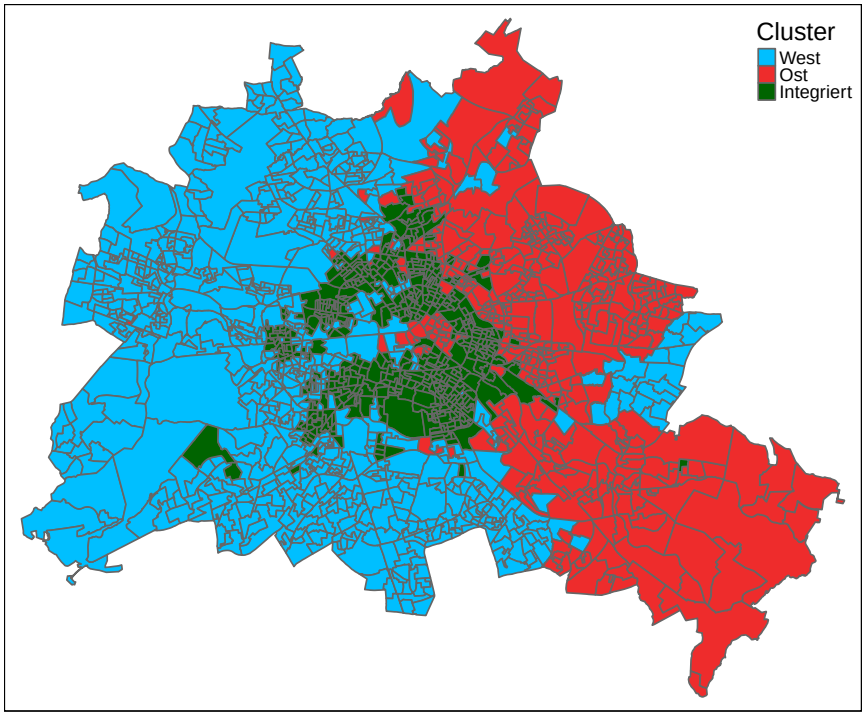

FIG. 6: Map of Berlin with districts color-coded to indicate their cluster-assignment from Fig. 3.

We then consider the question: does this pattern conform to any observable feature of the city? Fig. 8 shows the same map overlayed with the subway plan. Does support for the Green party coincide with the "Ring bahn" –perhaps their supports live near S-bahn stations and take public transit rather than drive cars? perhaps Green voters are younger, more interested in environmental issues, less invested in historical divisions, and more likely to be renting apartments near the inner city rather than owning homes in the surrounding neighbourhoods?

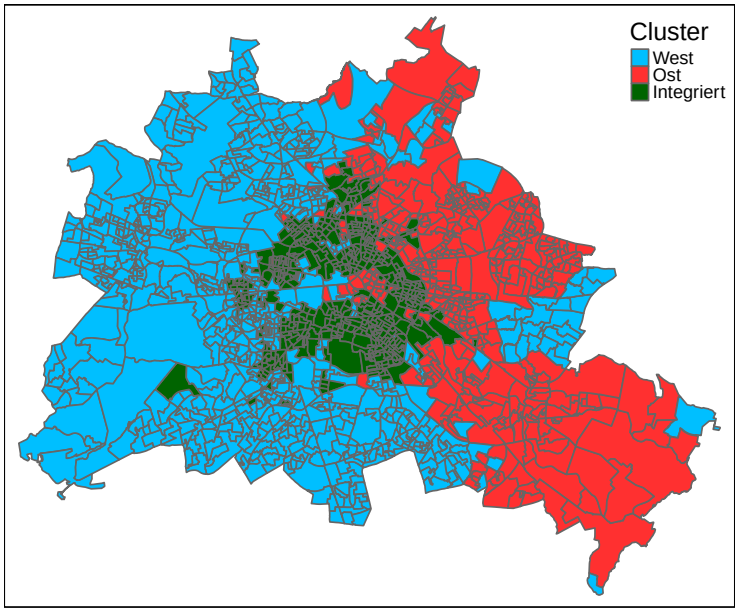

FIG. 7: Same map but using the results from the 1rst vote (i.e. based on candidate) -the results are almost identical.

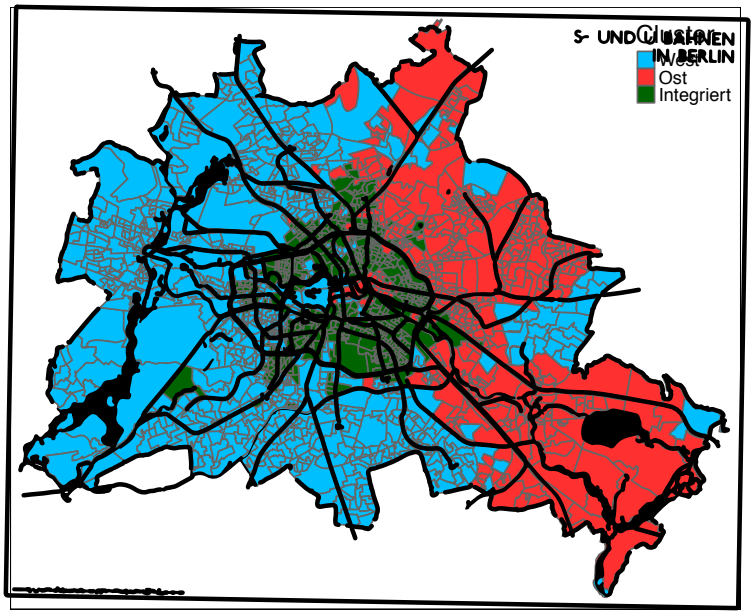

FIG. 8: The same data from Fig. 7, but with the U-bahn and S-bahn plan overlayed (unfortunately, the downloaded train lines that were obtained suffered from rendering issues in the resulting graph.) What other factors of the city might be correlated with this structure?

III. WHAT'S ACTUALLY CAUSING THIS?

We do not mean to imply that public transit is causing integration. Rather, this feature is symptomatic of a variety of other cultural and sociological factors: population density being one of the more obvious covariates. One might also consider, for example, the degree of immigration in a community (immigrants, presumably, being less attached to such a historical national division.) Fig. 9, indeed, shows some correlation with this trend, in areas such as Neuköln and Moabit, however, the above trend is apparently robust through the low-immigrant areas of PrenzlauerBerg and Wedding, while the latter appears alone in Charlottenberg.

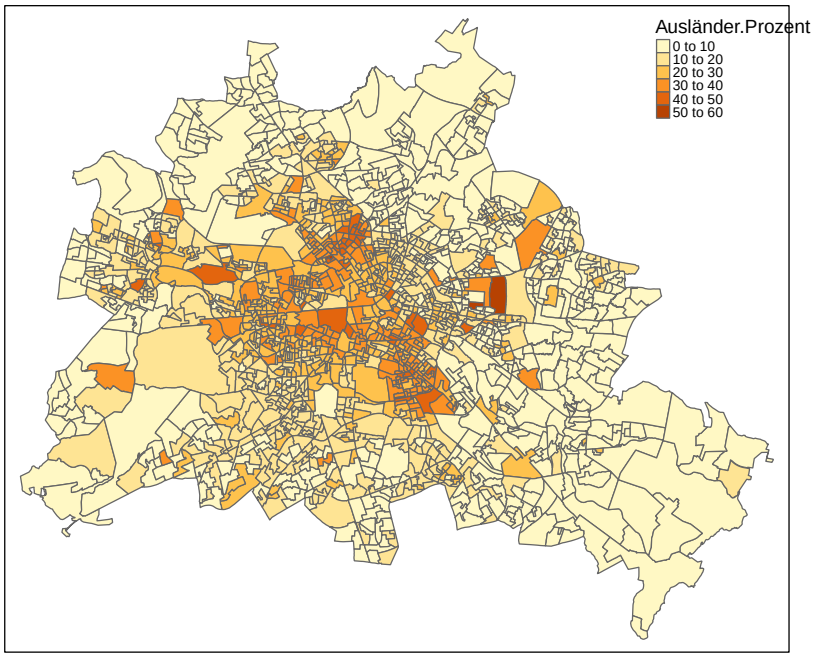

FIG. 9: Percentage of the population immigrating from outside of Germany.

We can also plot the age distributions: age ranges betwen 18-25, 45-60, and above 70 do not reveal much structure, however Fig. 10 seems to show relatively high voter percentages among the age group 25-35, while Fig. 11, conversely, shows and anti-correlation among the age group 60-70.

This would seem to paint a picture of integrated German society: Young and urbanized, but no strong correlation with immigration.

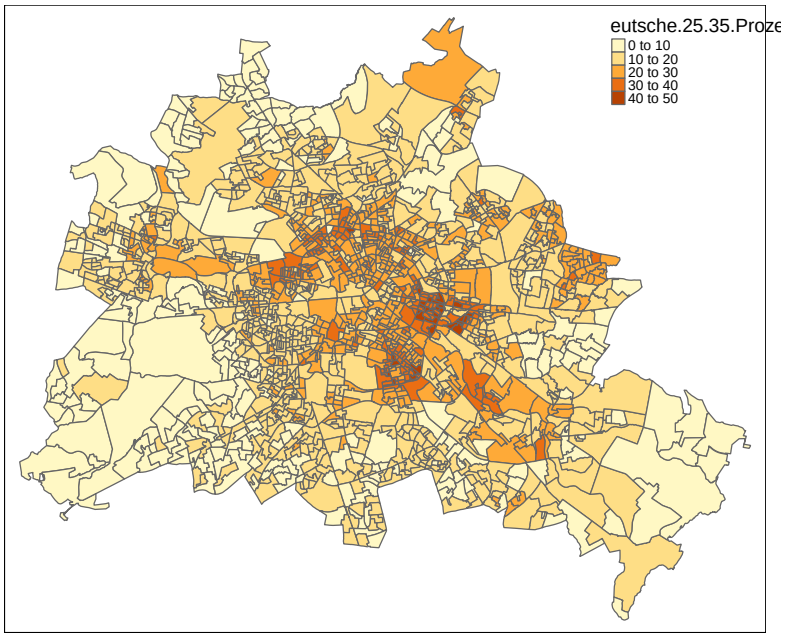

FIG. 10: Percentage of a ridings voters between the ages of 25 and 35: positive correlation with the Green 'integrated' ridings is seen qualitatively calculate p-values and add them here.

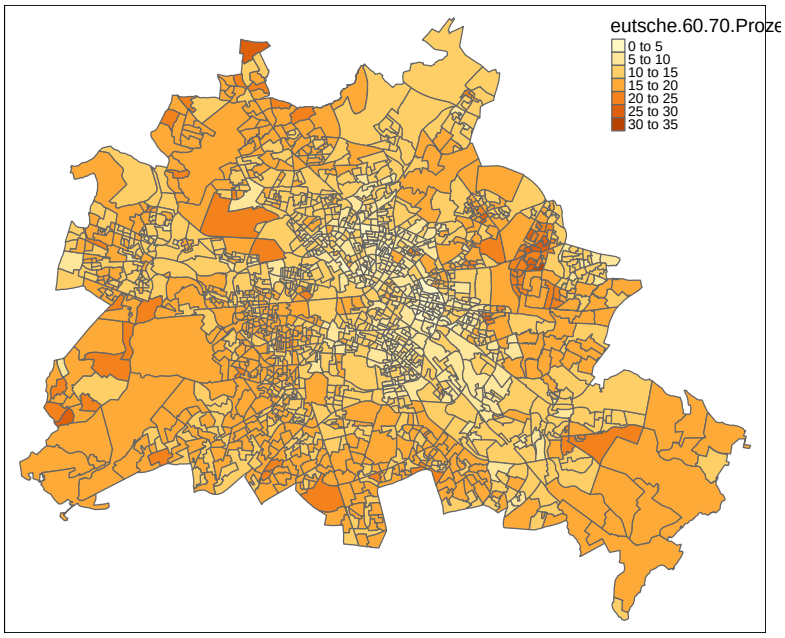

FIG. 11: Percentage of a ridings voters between the ages of 60 and 70: Anticorrelation with the Green-dominated integrated ridings is seen qualitatively, again: quantify this..

Other observations that can be made from this data:

- The Eastern cluster tends to be more politically polarized, with both Die Linke and AfD drawing major support. On top of this, the sum of major party support among the West cluster was 99.2%, while in the East cluster it was only 97.3% meaning that in the East cluster, citizens were more than 3.5 times as likely to vote for Fringe parties (*in addition* to their support for more radical parties such as AfD or die Linke) [3]

- Look for outliers to the clustering trend: e.g. that blue piece on the right that's physically east, but strongly ideologically 'West', is that the same as the 'tail' of red points at the bottom of Fig. 1?

- SPD support seems approximately uniform across the region -why are they only party whose support is independent of region?

- Why was Die Grüne so much higher in the integrated cluster? Does concern for the environment seems to transcend historical divisions?

IV. WHAT CAN BE EXTENDED TO THE NATIONAL LEVEL?

I have not yet processed geo-data of the bezirks at the national level, though the same trends may well scale up at national level. However, the same cluster analysis can be done on previous years' national elections to look for the same trends; seeing how demographics have shifted since the last election might provide some predictive power to subsequent election.

I'd also like to extrapolate correlation on cluster support with turnout rates -this would be particularly useful for party strategies: if the parties know what groups of people are supporting them, then they'd be interested to know where their efforts to get people to actually show up and vote are successful, and where they might need greater outreach. Before doing this, however, I thought it would be time-efficient to ask whether there is any interest in such results.

This, however, is a hobby project that I maintain occasionally alongside a full-time job. Hence, I have to be realistic about how much further time I can put in. If anyone would

[3] It may be worthwhile to check the distribution of "invalid" votes, or people who abstained from voting at all -perhaps this cluster correlates with turnout?.

be interested in collaborating on this and sharing the work or taking it further and doing something more interesting with it please let me know and I'd be happy to hand over the code or run it for you (talk to me and I'm sure we can agree to some fair attribution).
