Geography in Election Forensics[*]

Walter R. Mebane, Jr.[†] Kirill Kalinin[‡]

August 26, 2014

[*] Prepared for presentation at the 2014 Annual Meeting of the American Political Science Association, Washington, DC, August 28-31, 2014. Thanks to Joe Klaver and Jonathan Wall for assistance.

[†] Professor, Department of Political Science and Department of Statistics, University of Michigan, Haven Hall, Ann Arbor, MI 48109-1045 (E-mail: wmebane@umich.edu).

[‡] Ph.D. candidate, Department of Political Science, University of Michigan, Haven Hall, Ann Arbor, MI 48109-1045 (E-mail: kkalinin@umich.edu).

Abstract

We begin to investigate how geography affects the ability to draw inferences–based on reported vote counts–about election frauds or about other election features such as voters' strategies. Can we localize places where anomalies occur? It can be important to take into account geographic features such as local jurisdictional boundaries, but beyond that is proximity in physical space important? Should physically neighboring precincts be analyzed as having similar probabilities of frauds or of voters acting strategically, say, independent of whatever other covariates may be used in some model? We begin to study these concerns using data from several countries (Germany, Mexico, Russia).

Introduction

To consider geography in relation to election forensics can be interesting and important in a number of ways. Minimally one may wish to locate forensic statistics within existing political jurisdictions—say legislative districts—and observe how the statistics vary across geographic space. Mebane, Egami, Klaver and Wall (2014) has examples of that using data from Germany and Mexico. More generally one may wish to answer the question, "where are the anomalies?" (or "where are the frauds?"), with the idea that locations may not be constrained by predefined boundaries. For example, given precinct or polling station data, it may be that political strategies uniformly affect votes cast at all the precincts in particular legislative districts, or strategic considerations may span communities that reach across or lie only partly within individual districts. Fraudulent activities may or may not have reach that is coterminous with predefined jurisdictional boundaries. If polling station data is geolocated and even more if precinct boundaries are known, one might like to map measures of anomalous behavior and see how well they coincide with known juridictions. Beyond this it may be interesting to build geographic structure into models of behavior that relate to election strategies or frauds. For example, do the probabilities that petitions are submitted to nullify particular ballot boxes in Mexico (see Mebane et al. 2014) depend only on features of the instant ballot boxes or do they depend on what's happening in locations nearby.

One rationale for taking geographic space into account is the idea that there are unobserved variables and processes that are associated with spatial location. This idea motivates using geographic weights in statistical models: the statistic computed for each observation is a weighted sum both of the observation's own value and of the values of neighboring observations (e.g. McMillen and McDonald 2004; LeSage 2004). Such weights are usually functions of the distances between pairs of observations. Weights are nonnegative and are positive only for a set of an observation's nearest neighbors.

Given a variable that measures election anomalies in individual observations, we can try to detect clusters of observations that all exhibit the anomaly. One way to do this is to use a spatial scan statistic (Glaz and Balakrishnan 1999): one searches a map using a fixed size scanning “window” and checks whether what is found inside the window is significantly different from what is found outside the window. A Monte Carlo resampling method can be used to assess significance (Kulldorf 1997).

Some electoral data match the simplest case in which the vote count observations are all of the same type and the geometry has, in effect, the topology of a compact subset of two-dimensional Euclidean space. All data may be available for in-person voting precincts, for example, and we may have shapefiles or geolocations for the precincts, all of which are distinct. In such simple cases both distance and “neighbor” are straightforward to define. Distance is Euclidean distance and neighbor means “adjacent.”

But electoral data can have more complicated structure. Precincts may be a mix of types. For example, there can be both in-person precincts and mail precincts, and the boundaries of each mail precinct may encompass many in-person precincts. Voters are free to use either type of precinct to vote. Distances between precincts of the same type are well defined, but the topology is more complicated. Are nonadjacent in-person precincts that are included in the same mail precinct neighbors? Is an in-person precinct a neighbor of the mail precinct that contains it?

We briefly present three cases in which election forensics encounters geography. First we use data from the 2012 presidential election in Russia to illustrate use of a spatial scan statistic. In this case we have geolocated polling stations and a statistic that Mebane (2013*b*) identifies as being strongly associated with election frauds in Russia under Putin. Next we use data from the 2012 election in Mexico to illustrate use of geographic weights to compute spatially smoothed means. In this case we have sección (precinct) shapefile information, as well as shapefile information for other jurisdictions such as localities. We use locality shapefile information to distinguish urban from rural precincts, which we treat differently. Last we use data from the 2009 election in Germany, again to illustrate use of geographic weights. In this case we have precinct shapefiles only for twenty-two cities while for the rest of the country we know locations and boundaries only for Gemeinden (communities). Also in Germany we encounter the complication of having both in-person and mail precincts. In both Mexico and Germany we focus on the vote counts' second significant digits, which Mebane (2013*a*) argues strongly relate to voters' strategic behavior but which Pericchi and Torres (2011) claim are indicators of election frauds. In Germany we also consider another statistic that is often claimed (e.g. Bawn 1999) to indicate strategic voting.

Methods

The spatial scan methods we use are implemented in `SaTScan` (Kulldorf and Information Management Services, Inc. 2009). In particular, because the variable of interest in the Russian case is binary, we use the Bernoulli model (Kulldorf 1997).

To compute geographic weights we use Euclidean distances δ_{ih}, where i and h index observations, in the tri-cube function from Cleveland and Devlin (1988):

$$\omega_{ih} = \left[1 - \left(\frac{\delta_{ih}}{d_i}\right)^3\right]^3 I(\delta_{ih} < d_i) \tag{1}$$

where $I(\cdot)$ is the indicator function and d_i is the maximum distance for which the weight is positive (McMillen and McDonald 1997). McMillen and McDonald (2004) use a simple nearest-neighbor rule to determine d_i, but we use a variation. In Mexico we distinguish rural from urban precincts. For rural precincts "d_i is the distance of the qth nearest observation to i," as in McMillen and McDonald (2004, 227), except that only other rural precincts are included. Any urban precinct that is closer than d_i to rural observation i is also included. For urban precincts we use higher order neighbor sets to determine d_i: d_i is the distance to i of the furthest observation among the first through kth higher order neighbor set originating at i.[1] Any rural precinct that is closer than d_i to urban observation i is also included. In Germany we use a similar procedure with nearest neighbors and q determining d_i among communities and higher order neighbor sets and k determining d_i among precincts.[2] We use cross validation (McMillen and McDonald 2004, 238) to choose "bandwidths" q and k.

In Germany we compute estimates only for in-person precinct (*Urnenwahlbezirk*) or in-person community observations. Each mail precinct (*Briefwahlbezirk*) or mail community in Germany is included with a positive weight ω_{ih} for in-person observation i if any in-person observation that has positive weight is encompassed by the mail precinct or mail community. For mail precincts (in Berlin, Bremen, Bremerhaven, Hamburg and Stuttgart) we have shapefiles, but in all other cases the locations of the mail communities we identify are computed as the mean of the locations of the in-person precincts or communities they include. In cases where δ_{ih} for an included mail community is greater than d_i, we reset d_i to equal that mail community's δ_{ih}.

Scan Statistics in Russia

Kalinin and Mebane (2011) explain the predominance of turnout percentages that end in 0 or 5 by referring to the signaling strategies of Russian governors in the 2000s. According to the theory, the governors use rounded percentages of turnout as an easy and readily detected way to report their loyalty to superiors, which loyalty is in turn rewarded with intergovermental transfers in the post-electoral period. Mebane (2013*b*) shows that, in recent presidential elections, among precincts where Putin has more than sixty-five percent of the vote the absolute number of votes for Putin is higher in precincts where the vote percentage for Putin ends in 0 or 5. So whether Putin's vote percentage ends in 0 or 5 is a

[1] We use the function `poly2nb` in the `spdep` package of **R** (R Development Core Team 2005) to compute most of the first neighbors. For Germany we add some first neighbors uniquely identified using ArcGIS to the first neighbors identified by `poly2nb`. Islands are manually related to mainland observations based on inspection and judgments. We use `nblag` in the `spdep` package to compute the neighbor lags that identify the higher order neighbors.

[2] A further complication in Germany concerns mail precincts, discussed in the next paragraph.

plausible marker for election frauds in recent Russian elections. For the 2012 presidential election we define a precinct-level variable that has value one if Putin's vote percentage in the precinct ends in 0 or 5 and has value zero otherwise.

Geospatial analysis using scan methods helps detect geographic hotspots for the signaling strategies. In order to accomplish this task we use precinct-level data from 2012 presidential election. We use a Bernoulli model with data on geographical locations of precincts obtained from the GIS-Lab's website (`http://gis-lab.info`). We are able to match voting data for 91,067 precincts to precinct location information.[3]

Figure **??** shows that in 2012 several clusters of precincts exhibit strong signaling patterns. The figure shows outlines for regions in the western part of Russia. Red circles show cluster locations.[4] Regions are tinted blue if they contain a cluster. Ten clusters are identified, but only three are statistically distinctive according to Monte Carlo p-values: the clusters in Chechnya and Dagestan (including 352 precincts), in St. Petersburg (25 precincts) and in Bashkortastan (35 precincts). Two other republics have identifiable clusters but the differences between the clusters and the rest of Russia are not statistically significant: Tatarstan and Mordovia. Finding statistically significant clusters in Chechnya, Dagestan and Bashkortastan supports the fact that fraud is apparent in Republics, where the governors are most capable of mobilizing their regional "political machines" to provide electoral support to the national ruling elites in specific localities. Some Russian oblasts besides St. Petersburg also have identifiable clusters, but only the St. Petersburg cluster is statistically significant. These are Volgogradskaya oblast (eastern rural area), Kostromskaya (northern area), Saratovskaya oblast (south-west urban area), Samarskaya oblast (northern rural area) and Voronezhskaya (northern urban area). The displayed results suggest that signaling patterns can be rooted within specific localities of the regions.

*** Figure **??** about here ***

[3] Precinct-level data downloaded from sites connected to Central Election Commission of the Russian Federation (2013) have data for 95,415 precincts. Precincts that could not be geolocated are special precincts containing votes cast outside of Russia.

[4] Circle size represents the geographic area spanned by the precincts in each cluster.

Bandwidths for Geographically Weighted Means in Mexico and Germany

We expect that the bandwidths chosen for different kinds of variables in Mexico and in Germany may meaningfully relate to what the variables represent. We examine two kinds of variables: simple vote proportions; and the second digits of vote counts. While we have no specific expectations for what bandwidths may be for vote proportion means, we expect relatively large bandwidths for the second digit means. If as Mebane (2013*a*) argues the second digit means respond to voters' strategic behavior, then all voters in a single district ought to have strategic incentives with the same structure: if voters in each district coordinate through rational expectations as in the wasted vote model of Cox (1994, 1997), then all voters in the district should be focused on the same expected winner and expected first loser; given a "Duvergerian" equilibrium (Cox 1994), only two of the parties competing in each district should be receiving strategically switched votes. So if the second digits are sensitive to such strategic behavior, all or most of the precincts in each district should have the same mean. Mebane (2013*a*) demonstrates that the district means depend on the margins between the leading parties and the second losing parties, in line with (Cox 1994), but that analysis presumes district-based bundling of precincts. If in fact the second-digit means tend to vary over districts, then using fairly wide bandwidths will best tend to support mean estimates that vary over districts without needing to prespecify district grouping.

Pericchi and Torres (2011) might argue that likewise incentives to commit election fraud also have district-wide scope, so the occurrence of relatively wide bandwidths is not per se evidence that vote counts' second digits are better measures of strategy than they are measures of frauds. To decide whether strategies or frauds are in play—or both—we need to draw in other indicators of strategic or fraudulent behavior. In this paper we will do a tiny bit of that in relation to Germany.

Table 1 reports the bandwidths q and k selected for parties in legislative elections in

Mexico and Germany.[5] Bandwidths are selected independently for each variable of interest. For Mexico we report results only for parties identified using their place in each district's election results ("winner," "second" and "third"). For Germany we report results for parties identified by finishing placement in the *Erststimmen* and also by name in both the *Erststimmen* and the *Zweitstimmen*.[6] When simple vote proportions are estimated, both the q and k bandwidth values are much smaller than when the vote counts' second digits are estimated.

*** Table 1 about here ***

The higher values of the higher order neighbor set bandwidth k imply that many precincts are used to estimate the mean for each precinct. While q directly gives the number of nearest neighbors being used, the number of precincts involved for each value of k depends on the intricacies of precinct boundaries. Figure **??** reports these precinct counts for several of the bandwidth k values that appear in Table 1. The empirical densities in Figure **??** show the number of precincts contained in the first through kth higher order neighbor set originating at each i. For Germany the counts include both precincts (in the cities for which we know precinct boundaries) and communities. For $k = 1$ in Mexico and for $k = 2$ in Germany, the numbers of precincts and communities are small, often single digits in Mexico and typically about 20 in Germany. As k rises the number of included precincts or communities rises rapidly. For $k = 7$ in Germany there are typically just under 400 precincts and communities. For $k = 12$ in Mexico and for $k = 15$ in Germany, the numbers rise to typically between 1,000 or 2,000. For $k = 20$ in Mexico the numbers rise to typically 3,000 to 4,000. In both countries the higher values of k are ample to span entire election districts.

*** Figure **??** about here ***

[5] Data source information for each country appears in subsequent subsections.

[6] In German Bundestag elections each voter simultaneously casts two votes. The *Erststimme* is a vote for a plurality rule winner in an election district, and the *Zweitstimme* is a vote for a party using proportional representation rules.

The wasted-vote rationale analyzed by Cox (1994) does not apply to the German *Zweitstimmen*, but Germany's mixed system suggests that the *Zweitstimmen* should also be affected by strategic considerations. So-called "coalition threshold insurance" arises when some supporters of larger parties (CDU-CSU or SPD) cast their *Zweitstimmen* for one of the smaller parties (FDP or Green) in order to support that party's involvement in a coalition in the Bundestag. Evidence using survey-based preference measures suggests both kinds of strategy occur (Pappi and Thurner 2002; Gschwend 2007; Shikano, Herrmann and Thurner 2009).

Geographically Weighted Means in Mexico

We estimate geographically weighted means for several variables based on votes in the Mexico 2012 House of Deputies elections. Vote count data at the casilla (ballot box) level come from Instituto Federal Electoral (2012). We use data aggregated to the sección (precinct) level. There are 66,494 secciones. Sección shapefiles and other shapefiles originate from IFE and INEGI.[7]

Figure **??** shows the results of estimating second-digit means for winning, second-place and third-place candidates across all of Mexico. Let $\hat{j}_i$ denote the second-digit mean estimate for observation i. Colors reflect how the mean estimated for each sección relates to the mean of $\bar{j} = 4.187$ expected if the second digits are distributed according to Benford's Law.[8] Mebane (2013*a*) identifies the value $\bar{j}$ as the value that occurs when there is an essentially tied two-party election and there is no strategic behavior. Mebane (2013*a*) finds that district imbalances and strategic voting can both—in distinctive ways—produce second-digit means higher or lower than $\bar{j}$. In the simple case of three parties competing

[7] We obtained shapefiles via Valle-Jones (2013*b*,*a*), which provide scripts to download from `http://www.ife.org.mx`, `http://gaia.inegi.org.mx/` and `http://mapserver.inegi.org.mx/`. The electoral shapefiles needed to be cleaned up in various ways. In particular, sección shapes were originally not connected across Mexican state boundaries and states were slightly misaligned. Our production shapefile includes 3960 null observations (a few of which are visible) due to steps taken to unify the sección shapefile.

[8] White splotches reflect either imperfections in the shapefile (recall note 7) or precincts where the variable is missing. We do not estimate means to fill in missing data, although it is easily possible to do so.

with wasted-vote strategic activity, second-digit means near 4.35 often occur. Mebane (2013*a*) finds that other kinds of strategic behavior can produce even higher second-digit means. Strategic voting in multiparty situations can produce second-digit means somewhat less than $\bar{j}$. Mebane (2013*a*) finds that for parties that are stratgically abandoned, such as the third party in the simplest wasted-vote scenario, the second digit mean is often substantially less than $\bar{j}$. In Figure **??** the color green indicates $\hat{j}_i$ is very near $\bar{j}$, the color red indicates $\hat{j}_i > \bar{j}$ and the color blue indicates $\hat{j}_i < \bar{j}$.[9]

*** Figure **??** about here ***

Figure **??** presents ample evidence of widespread strategic behavior. In the map for winning parties there are ample reddish areas, a few blue tinged areas and some green areas. The map for second-place parties similarly has extensive reddish areas, some blue areas and some green areas. The map for third-place parties is mostly blue and green. These patterns are roughly what we expect if voters are using wasted-vote strategies and if the vote counts' second digits are sensitive to the strategic behavior.

Using the mean estimates we can focus on any local area comprised of secciones. We consider three areas. Figure **??** displays second-digit means for Distrito Federal surrounded by parts of neighboring states. Black lines in the figure show district boundaries while white lines outline secciones. Distrito Federal occupies the roughly heart-shaped area in the middle of each subfigure. For winners the dominant hue is reddish, for second-place parties the color is reddish but with more green, and for third-place parties the color is bluish with some green. In Figure **??** we look at the whole state of Mexico, which wraps aroung Distrito Federal. For winners the predominant hue is reddish, for second-place parties there is a mix of red, blue and green, and for third-place mostly blue with some green. The differences between the winner's and the third-place parties's means are largely compatible with what wasted-vote strategy implies—if the second digits are sensitive to the

[9] To be specific, we transform $\hat{j}_i$ using $\mathfrak{j}_i = \text{logistic}(\hat{j}_i - \bar{j})$ and then choose the color for observation i using the **R** commands `za<-abs(z-.5)^(1/5); colstr(z,(1-za),(1-z));`, where `z` is $\mathfrak{j}_i$ and the respective arguments of `colstr()` fix the intensity of the red, green and blue components.

strategic behavior—but the differences between the second-place and third-place parties' means are somewhat less compatible.

*** Figures ?? and ?? about here ***

The third area we consider is Oaxaca, displayed in Figure ?? (with a bit of Guerrero to the west, Puebla and Veracruz to the north and Chiapas to the east). Figure ?? shows predominantly reddish and green colors for winners and second-place parties but predominantly bluish and green colors for third-place parties. The patterns are compatible with voters using wasted-vote strategies.

*** Figure ?? about here ***

Geographically Weighted Means in Germany

We estimate geographically weighted means for the German 2009 Bundestag election. Vote count data at the polling station level come from Bundeswahlleiter (2011). Shapefiles are produced by combining December 2009 community (Gemeinden) shapefiles from Bundesamt für Kartographie und Geodäsie (2013) with precinct shapefiles obtained from several cities.[10] Cities for which we have 2009 *Urnenwahl* precinct shapefiles are Berlin, Braunschweig, Bremen (including Bremerhaven), Dortmund, Freiburg, Hamburg, Hamm, Hannover, Herne, Kassel, Kiel, Köln, Leipzig, Magdeburg, Mainz, Mönchengladbach, München, Münster, Regensburg, Rostock, Stuttgart and Wuppertal.[11] Cities for which we have 2009 *Briefwahl* precinct shapefiles are Berlin, Bremen (including Bremerhaven), Hamburg and Stuttgart. Most of the shapefiles require editing in order to fit together and

[10] The specific Gemeinden shapefile is in `vg250_20091231.utm32s.shape.ebenen.zip`.

[11] Cities that supplied data about 2009 geographies that were not usable for one reason or another are Bonn, Duisberg, Frankfurt, Ludwigshafen and Saarbrücken. Other cities that provided geographic information (usually shapefiles) but not for 2009 include Düsseldorf, Erfurt, Gelsenkirchen, Karlsruhe, Leverkusen, Lübeck, Solingen and Wiesbaden. We obtained the data first by emailing the *Bundeswahlleiter*, which we learned does not maintain precinct shapefiles, then hundreds of *Kreiswahlleitungen* and then the election offices, statistics offices and cartography/surveying/geodata offices of each of the twenty-five most populous German municipalities. Many of the *Kreiswahlleitungen* responded that their communities are too small to have precinct shapefiles, and in fact 5,503 Gemeinden have only a single in-person precinct.

to allow the vote count data to be merged with the precinct and community boundary information. For the analysis in this paper, data from all *Urnenwahl* for which we lack precinct shapefiles are summed by community to produce community-level counts. In final form our shapefile includes 11,989 in-person community observations and 8,817 in-person precinct observations. Data from all *Briefwahl* for which we lack mail precinct shapefiles are summed by mail community to produce mail community counts. *Gemeindefrei* areas (areas with no residents, like parks and forests) are assigned counts of zero for all vote variables. Some water features are removed.[12] The `RAU_RS` code is a basis for associating in-person precincts and communities with their mail communities, although matching is not an entirely straightforward process.[13] All together we have 1,022 mail community and 4,022 mail precinct observations.

In-person precincts and communities have two-dimensional Euclidean geometry and topology, but the topology of mail precincts and communities is special.[14] Each mail precinct or community is an immediate neighbor of the in-person observation that it encompasses, although the mail observation is not counted as a nearest neighbor or as a member of any higher order neighbor set, and each mail observation is included with positive weight at most once when computing the mean for each in-person observation. Every mail precinct or community h that encompasses an included neighbor of observation i has positive weight when computing $\hat{j}_i$, unless $\delta_{ih} = d_i$. In-person precincts or communities that are not otherwise among the q nearest neighbors or the k higher order neighbor sets for i do not get positive weight for $\hat{j}_i$ just because they are encompassed by a mail precinct or community that encompasses i or an included neighbor of i.

[12] We delete all features where attibute GF = 2, which removes 83 polygons from the Gemeinden shapefile.

[13] The 2009 voting data *Briefwahlbezirke* cover 5,601 distinct `RAU_RS` codes. All but 832 of the `RAU_RS` codes occur in the Gemeinden shapefiles. Of those 832 all but three `RAU_RS` codes end in '999'. 502 of the unmatched `RAU_RS` codes match if only the first nine digits are used (all but the "Gemeinde" code). The remaining `RAU_RS` codes match when only the first nine digits (comprising the "Land," "Regierungsbezirk" and "Kreis" codes) are used.

[14] The locations of mail communities are computed as the mean of the locations of the in-person precincts or communities they encompass, which is the set of in-person precincts or communities with matching full or partial `RAU_RS` codes.

A variable often used to study strategic voting in Germany is the difference in votes received by a party in the *Zweitstimmen* compared to the *Erststimmen* (e.g. Bawn 1999). The leading parties in a district's *Erststimmen* can be expected to receive a smaller proportion of the *Zweitstimmen* because the wasted-vote strategy that gains them votes in the plurality rule *Erststimmen* has no effect in the proportional representation rule *Zweitstimmen*. A small party may also receive a higher proportion of votes in the *Zweitstimmen* than in the *Erststimmen* due to threshold insurance strategies.

Estimates for the geographically weighted mean difference between *Zweitstimmen* and *Erststimmen* proportions for each precinct or community are broadly compatible with those strategic stories. Figure **??** shows the means for CDU-CSU and for SPD, the two largest parties. Red indicates that the *Zweitstimmen* proportion is greater than the *Erststimmen* proportion, blue indicates that the *Erststimmen* proportion is greater than the *Zweitstimmen* proportion, and green indicates the two proportions are the same.[15] Figures for both parties are predominantly blue—more *Erststimmen* than *Zweitstimmen*—with occasional red areas and a very few patches of green. For small parties and particularly for FDP the situation is reversed. In Figure **??** red dominates the map for FDP: more *Zweitstimmen* than *Erststimmen*. Red dominates the map for Green in the west, but in the south and east much is blue. For Die Linke a generally reddish color dominates across much of the country, but in the east there is extensive blue.

*** Figures **??** and **??** about here ***

Figure **??** shows $\hat{\jmath}_i$ for *Erststimme* winning, second-place and third-place candidates. For winners there are wide expanses of red areas and blue areas, separated by thin areas of green. For second-place parties there are also wide areas of red and of blue with perhaps slightly less red than appears for winners and also more green. These patterns might suggest that strategic voting using wasted-vote strategies is widespread. The fact that the

[15] Colors are determined using the scheme described in note 9 except with $\mathrm{j}_i = \text{logistic}(100\hat{\jmath}_i)$. White splotches in Figure **??** reflect *Gemeindefrei* areas. In other figures for Germany white splotches may come from either *Gemeindefrei* areas or observations where the variable is missing.

map for third-place parties has as much red as does the map for second-place parties is a reason to hesitate about that interpretation. Perhaps these results do reflect wasted-vote strategies in play, but Germany's mixed system means that wasted-vote strategic behavior in the *Erststimmen* is not quite as simple as described in the Duvergerian equilibria of Cox (1994).

*** Figure **??** about here ***

Figure **??** shows $\hat{j}_i$ for *Erststimmen* second digits for parties by name instead of by finishing position. Results for four parties—CDU-CSU, FDP, SPD and Green—are in Figure **??** (results for Die Linke are in Figure **??** below). All four parties show areas of red, blue or green. The array of colors is comparable for the *Zweitstimmen* second digits, shown in Figure **??**, although the same places only occasionally have similar second digit mean values across voting rule types. Colors are the same in both the *Erststimmen* and the *Zweitstimmen*, for instance, for CDU-CSU in the west of Germany in the vicinity of Dortmund and in the northeast in Mecklenburg-Vorpommern. Colors are also similar across voting rules in pretty much the same areas for SPD.

*** Figures **??** and **??** about here ***

For Die Linke second digit means in the same locations are similar across voting rules in much of Germany. The similarities in color between *Erststimmen* and *Zweitstimmen*, in Figure **??**, is remarkable, especially in contrast to the differences across voting rules observed for the other parties in Figures **??** and **??**. The similarities for Die Linke are especially pronounced in the northern part of the country. A brief way to explain why Die Linke differs from the other parties is that Die Linke is not involved in governing coalition considerations as the other parties are. So perhaps the task for voters regarding Die Linke is a bit simpler than for some of the other parties.

*** Figure **??** about here ***

Interesting contrasts between jurisdictions become apparent when we focus on some particular areas. Figure **??** focuses on $\hat{j}_i$ for second digits of votes for *Erststimmen* winners, second- and third-place finishers in Berlin. The contrast betwen $\hat{j}_i$ for winners in Berlin compared to winners in Brandenburg around Berlin is remarkable. Inside Berlin winners' colors are all the same shade of blue, while in Brandenburg winners have the same reddish shade. For second-place parties the contrast between Berlin and Brandenburg is not as sharp and a greenish color dominates, although the color in districts in central Berlin is strongly blue. For third-place parties we have blue in Brandenburg around Berlin and greenish blue inside Berlin. There are strong signs of relative homogeneity in strategic behavior inside districts in Berlin and in at least the parts of the districts we can see in Brandenburg.

*** Figure **??** about here ***

The impression of strategic homogeneity—or at least homogeneity in colors—within districts based on $\hat{j}_i$ for *Erststimmen* second digits is mostly confirmed when we pull back and look at all of Brandenburg. In Figure **??** the districts around Berlin are strongly, although not perfectly, internally homogeneous in their colors. For winner the colors are reddish immediately around Berlin and more blue further south and west. For second-place parties it's mostly greenish immediately around Berlin, with reds and blues further to the west and south. For third-place parties the color is blue immediately around Berlin.

*** Figure **??** about here ***

The last local area we examine is Bavaria. $\hat{j}_i$ for *Erststimmen* second digits in Bavaria are shown in Figure **??**. For both winners and second-place parties are reddish for districts in and around München and Stuttgart. For third-place parties the color around München is bluish while the color around Stuttgart is reddish. Not only in these districts but generally in and around Bavaria the impression is of predominantly homogeneity in color within particular districts, be the color red, blue or green. Such patterns are compatible with a conclusion that the vote counts' second digits are sensitive to strategies. This pattern occurs even though, unlike in Mebane (2013*a*), the vote counts in point are not uniformly precinct vote counts.

*** Figure **??** about here ***

Discussion

We find suggestive patterns when bringing geography into election forensics in three countries, even when the statistics being used are simple means. The results point to plausible clusters of election frauds in Russia. For Mexico and Germany the results give reasonable signs that voters are acting strategically. The interpretations in all these cases are of course merely suggestive, but they are at least suggestive.

While the process of assembling geographic information and merging it with electoral data can be arduous, the prospects for high payoffs from doing so are good. The general point is that geographic information may allow strategies, frauds and other anomalies to be located without having to rely on preexisting political jurisdictions. Among the tools to be developed are ways to test formally whether existing jurisdiction boundaries are in fact supporting geographic structure that may appear in the data. Monte Carlo resampling schemes like those used with scan statistics seems a likely way to proceed, and indeed Monte Carlo methods are well developed for that purpose in that literature. A question is the feasibility of performing such tests with statistics more complicated than the mean. Another direction is to build geography into election fraud estimation methods such as the finite mixture version of the Klimek, Yegorov, Hanel and Thurner (2012) model that Mebane et al. (2014) have begun to develop. Much remains to do.

References

Bawn, Kathleen. 1999. "Voter Responses to Electoral Complexity: Ticket Splitting, Rational Voters and Representation in the Federal Republic of Germany." *British Journal of Political Science* 29(3):487–505.

Bundesamt für Kartographie und Geodäsie, Dienstleistungszentrum des. 2013. "Index of /auftrag1/archiv/vektor/vg250_ebenen." Accessed February 16, 2014. URL: `http://www.geodatenzentrum.de/auftrag1/archiv/vektor/vg250_ebenen/`.

Bundeswahlleiter, Der. 2011. *Wahl zum 15. Deutschen Bundestag am 22. September 2002.* Wiesbaden: im Auftrag der Herausgebergemeinschaft. Ergebnisse der Wahlbezirksstatistik: Die Wahlleiter des Bundes und der Länder, Auszugsweise Vervielfältigung und Verbreitung mit Quellenangaben gestattet.

Central Election Commission of the Russian Federation. 2013. "Elections and referendums." URL http://www.vybory.izbirkom.ru/.

Cleveland, William S. and Susan J. Devlin. 1988. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* 83(403):596–610.

Cox, Gary W. 1994. "Strategic Voting Equilibria Under the Single Nontransferable Vote." *American Political Science Review* 88:608–621.

Cox, Gary W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems.* New York: Cambridge University Press.

Glaz, Joseph and N. Balakrishnan. 1999. *Scan Statistics and Applications.* New York: Wiley.

Gschwend, Thomas. 2007. "Ticket-splitting and Strategic Voting under Mixed Electoral Rules: Evidence from Germany." *European Journal of Political Research* 46(1):1–23.

Instituto Federal Electoral. 2012. "Estadísticas y Resultados Electorales." Descarga de la base de datos de los Cómputos Distritales, file `datos_computos_2012_09072012_2015.zip`, URL: `http://www.ife.org.mx/portal/site/ifev2/Estadisticas_y_Resultados_Electorales/` (accessed October 19, 2012).

Kalinin, Kirill and Walter R. Mebane, Jr. 2011. "Understanding Electoral Frauds through Evolution of Russian Federalism: from "Bargaining Loyalty" to "Signaling Loyalty"." Paper presented at the 2011 Annual Meeting of the Midwest Political Science Association, Chicago, IL, March 31–April 2.

Klimek, Peter, Yuri Yegorov, Rudolf Hanel and Stefan Thurner. 2012. "Statistical Detection of Systematic Election Irregularities." *Proceedings of the National Academy of Sciences* 109:16469–16473.

Kulldorf, Martin. 1997. "A Spatial Scan Statistic." *Communications in Statistics* 26(6):1481–1496.

Kulldorf, Martin and Information Management Services, Inc. 2009. *SaTScanTM v8.0: Software for the Spatial and Space-time Scan Statistics.*
URL: *http://www.satscan.org/*

LeSage, James P. 2004. A Family of Geographically Weighted Regression Models. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer.

McMillen, Daniel P. and John F. McDonald. 1997. "A Nonparametric Analysis of Employment Density in a Polycentric City." *Journal of Regional Science* 37:591–612.

McMillen, Daniel P. and John F. McDonald. 2004. Locally Weighted Maximum Likelihood Estimation: Monte Carlo Evidence and an Application. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer.

Mebane, Jr., Walter R. 2013*a*. "Election Forensics: The Meanings of Precinct Vote Counts' Second Digits." Paper presented at the 2013 Summer Meeting of the Political Methodology Society, University of Virginia, July 18–20, 2013.

Mebane, Jr., Walter R. 2013*b*. "Using Vote Counts' Digits to Diagnose Strategies and Frauds: Russia." Paper presented at the 2013 Annual Meeting of the American Political Science Association, Chicago, August 29–September 1, 2013.

Mebane, Jr., Walter R., Naoki Egami, Joeseph Klaver and Jonathan Wall. 2014. "Positive Empirical Models of Election Fraud (that May Also Measure Voters' Strategic Behavior." Paper presented at the 2014 Summer Meeting of the Political Methodology Society, University of Georgia, July 24–26, 2014.

Pappi, Franz Urban and Paul W. Thurner. 2002. "Electoral Behaviour in a Two-vote System: Incentives for Ticket Splitting in German Bundestag Elections." *European Journal of Political Research* 41(2):207–232.

Pericchi, Luis Raúl and David Torres. 2011. "Quick Anomaly Detection by the Newcomb-Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela." *Statistical Science* 26(4):502–516.

R Development Core Team. 2005. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL: `http://www.R-project.org`.

Shikano, Susumu, Michael Herrmann and Paul W. Thurner. 2009. "Strategic Voting under Proportional Representation: Threshold Insurance in German Elections." *West European Politics* 32(3):634–656.

Valle-Jones, Diego. 2013*a*. "diegovalle/download-maps12." Accessed December 16, 2013. URL: `https://github.com/diegovalle/download-maps12`.

Valle-Jones, Diego. 2013*b*. "Download electoral shapefiles of Mexico." February 27, 2013. Accessed December 16, 2013. URL: `http://blog.diegovalle.net/2013/02/download-shapefiles-of-mexico.html`.

Table 1: Cross-validation Selected Bandwidths, Mexico 2012 and Germany 2009

	Mexico 2012					Germany 2009			
	Proportion		Second-digit			Proportion		Second-digit	
party	*q*	*k*	*q*	*k*	party	*q*	*k*	*q*	*k*
winner	8	1	832	12	winner	8	2	336	7
second	8	1	960	18	second	16	2	344	15
third	8	1	640	20	third	8	2	344	7
					E CDUCSU	8	2	336	8
					E SPD	8	2	168	6
					E FDP	16	2	344	20
					E Green	8	2	344	8
					E Die Linke	8	2	328	20
					Z CDUCSU	8	2	248	5
					Z SPD	8	2	256	8
					Z FDP	16	2	328	10
					Z Green	8	2	328	7
					Z Die Linke	8	2	336	20

Note: “winner,” “second” and “third” denote district winners in House of Deputies elections in Mexico and in Bundestag *Erststimmen* in Germany. Bandwidths for particular parties named in Germany are for either *Erststimmen* (“E”) or *Zweitstimmen* (“Z”).