Could the outcome of the 2016 US elections have been predicted from past voting patterns?

PMU Schmitz[1,2,3], JP Holloway[1], N Dudeni-Tlhone[1], B Ntlangu[4], R Koen[1]

1. CSIR Built Environment, Meiring Naude Rd, Brummeria, Pretoria, South Africa; pschmitz@csir.co.za, jholloway@csir.co.za, ndudenitlhone@csir.co.za, rkoen@csir.co.za

2. Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Lynnwood Rd, Hatfield, Pretoria, South Africa.

3. Fakultät für Vermessung, Informatik und Mathematik, Hochschule für Technik, Stuttgart, Schellingstrasse 24, D-70174, Stuttgart, Germany.

4. CSIR Modelling and Digital Science, Meiring Naude Rd, Brummeria, Pretoria, South Africa; bntlangu@csir.co.za

Abstract: In South Africa, a team of analysts has for some years been using statistical techniques to predict election outcomes during election nights in South Africa. The prediction method involves using statistical clusters based on past voting patterns to predict final election outcomes, using a small number of released vote counts. With the US presidential elections in November 2016 hitting the global media headlines during the time period directly after successful predictions were done for the South African elections, the team decided to investigate adapting their method to forecast the final outcome in the US elections. In particular, it was felt that the time zone differences between states would affect the time at which results are released and thereby provide a window of opportunity for doing election night prediction using only the early results from the eastern side of the US. Testing the method on the US presidential elections would have two advantages: it would determine whether the core methodology could be generalised, and whether it would work to include a stronger spatial element in the modelling, since the early results released would be spatially biased due to time zone differences. This paper presents a high-level view of the overall methodology and how it was adapted to predict the results of the US presidential elections. A discussion on the clustering of spatial units within the US is also provided and the spatial distribution of results together with the Electoral College prediction results from both a 'test-run' and the final 2016 presidential elections are given and analysed.

Keywords: Elections, clustering, predictions, counties

1. Introduction

In South Africa, a team of analysts has for some years been using statistical clustering techniques to predict election outcomes during election nights. This team represents a successful collaboration of computer scientists, GIS practitioners and statisticians who have developed a model to produce real-time updates to forecasts as the "live" information on vote counts is released.

The method involves using clusters based on past voting patterns to predict final election outcomes during election nights, once a sample of between 7 to 10 percent of voting districts have been declared, and has been used successfully in a number of South African elections as illustrated in Greben et al. (2005) and Greben et al. (2006). After the success achieved in the 2016 South African elections, in which the forecasts correctly predicted final outcomes which were not generally expected in the run-up to the elections, the question arose as to whether the method could be generally applicable to elections carried out in other countries.

With the US presidential elections in November 2016 hitting the global media headlines during the time period directly after the South African elections, the team decided to investigate adapting their method to forecast the final outcome in the US elections. It was felt that the modelling method could potentially utilise the fact that the voting results from states in earlier time zones are released before those in later time zones. South African voting districts all fall within one time zone, but the US voting precincts span 4 continental time zones, namely Eastern, Central, Mountain and Pacific Time Zone and two extra time zones, one for Alaska and one for Hawaii. Testing the method on the US presidential elections would have two advantages: it would determine whether the core methodology could be generalised, and whether it would work to include a stronger spatial element in the modelling, since the early results released would be spatially biased due to time zone differences.

2. Methodology

This section provides a summary of the South African-based election night forecasting model and how it was applied to the 2016 US presidential elections. The two main aspects discussed are the grouping of the voting population into segments with similar voting behaviour and the decision regarding the spatial units on which the forecasts would be based. As indicated in the previous section, this model was developed for the South African elections and has been applied during the 1999, 2004, 2009, 2014 general elections and the 2000, 2006, 2011, 2016 municipal elections. The mathematical details of this forecasting model can be obtained in (Greben et al., 2006).

2.1 Clustering of voting patterns

The core methodology of the South African election model involves fuzzy clustering of voting districts. Clusters are composed from voting patterns in a previous election period(s) so that new voting counts can be used in an "intelligent" way to forecast overall patterns in the new election.

The model assumes that voting behaviour is not random but has a statistically quantifiable pattern. This assumption implies that voting behaviour is influenced by political, socio-economic and demographic factors as well as past voting history. Another assumption is that changes in voting behaviour do not occur at random; hence, this model relies on clustering the electorate at the voting district (VD) level into segments with similar voting patterns. In this case, a voting district refers to the smallest spatial unit at which voting occurs and at which the voting results are released. The role of GIS is mainly to ensure that spatial changes in VDs are correctly mapped from one voting period to the next so that the forecasts maintain the correct information consistency regarding voting patterns.

The characteristics of the fuzzy clusters are such that each VD has membership in every cluster based on their previous voting profile (voting percentages per party). The use of "20" as the number of clusters for the model has been tested and validated and observed to be efficient in predicting the South African elections. Clustering also involves determining the appropriate measure of similarity - a distance measure in mathematical terms. In creating election clusters we use the Euclidean distance. The motivation for using the Euclidean distance as opposed to, for instance, using a standardised Euclidean distance, is that the Euclidean distance gives higher weighting to larger political parties, thereby giving emphasis to the voting patterns of larger political parties rather than smaller parties. Consequently, the forecasting of the election outcome for larger parties is expected to be better and more stable since these parties are well represented by the clusters. This has always been a desirable property since the performance of the larger parties is usually of greatest interest.

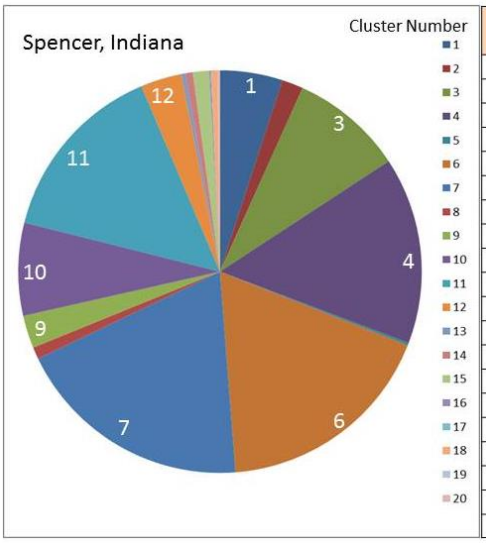
This model has, over a number of elections in South Africa, produced good predictions at various spatial or administrative levels (municipal, provincial and national scales) and has attracted a great deal of publicity among the public, the national broadcaster as well as political analysts. We therefore felt it would be appropriate to test this methodology elsewhere and in this case we questioned whether similar clusters could be compiled for the US - a much bigger country with a much higher voting population than in South Africa.

2.2 Spatial unit for clustering and forecasting the election outcome

The spatial unit at which clusters are formed has always been an important aspect in the prediction of the South African elections. In the South African context every registered voter belongs to a VD which in turn gets grouped proportionally to each of the resulting clusters to represent a voting tendency. The voting results are also verified and released at a VD level. Statistically speaking, the order in which the voting results are released has a pattern. For instance, election results from the urban areas tend to be released much earlier than those from the rural parts of the country. Most recently, however, smaller VDs (voting districts with few numbers of registered voters) tend to get declared sooner than the larger VDs. Also, in the case of disputes, complaints and the sharing of ballot papers across VDs, the South African Independent Electoral Commission (IEC) may withhold the release of the results until resolution or the verification process is complete. Therefore due to this level of bias in the order by which results get released, there is usually a considerable difference between the "scoreboard" results and the final election outcome. This affords the election forecasting model an opportunity to provide useful insights about what the initial results really mean.

It was therefore important for the team to determine an appropriate spatial resolution at which the US presidential elections should be modelled and the availability of data was a key factor to that decision. Initially attempts were made to source US data at a voting precinct level, which is equivalent to the South African voting district, but ultimately this was not a feasible option due to a number of reasons. Firstly, data on previous voting results and registered voters were not consistently available across all states at precinct level; secondly, the precincts are affected by spatial changes from one election period to another and the corresponding GIS data for these changes were not available to the team, and thirdly, a website containing a live release of voting precinct data was not available for the election night predictions. Consequently, since county boundaries remained fairly constant over election periods and election results per county could be obtained both historically and on the election night, the counties were selected as the smallest spatial unit for clustering. Data on registered voters per county remained a challenge since not all states had up-to-date registered voter counts available before the election day. However, estimates for the missing states could be obtained by using various other sources, such as previous elections.

Figure 1 shows an example for Spencer County in Indiana which had a mixed membership amongst the clusters (where clusters are based on 2012 voting patterns), while Figure 2 shows the clustering memberships for some counties in Indiana. It is clear from Figure 2 that, apart from Spencer County, most counties have a strong membership in either one or two clusters. In addition, the typical spatial nearest-neighbour principle is generally not always evident in the election clusters, as illustrated in Figure 2.



Cluster Number	Percent Democrat	Percent Republican	Percent Other	Membership
1	52.1	46	1.8	0.05
2	57.3	40.8	1.9	0.02
3	49.3	48.8	2	0.09
4	46.3	51.8	1.9	0.15
5	69.9	28.3	1.8	0.00
6	40.7	57.4	1.9	0.18
7	43.5	54.5	1.9	0.19
8	60.3	37.9	1.8	0.01
9	55.1	42.8	2	0.03
10	34.3	63.7	2	0.07
11	37.8	60.1	2.1	0.15
12	30.5	67.6	2	0.03
13	66.2	32.1	1.7	0.00
14	62.9	35.2	1.9	0.01
15	26.2	72	1.9	0.01
16	74.1	24.2	1.7	0.00
17	78.9	19.7	1.4	0.00
18	21.1	77.1	1.8	0.01
19	87.3	11.9	0.9	0.00
20	13.6	84.8	1.5	0.00

Fig. 1. Example of a mixed cluster membership – Election cluster membership of Spencer County in Indiana.

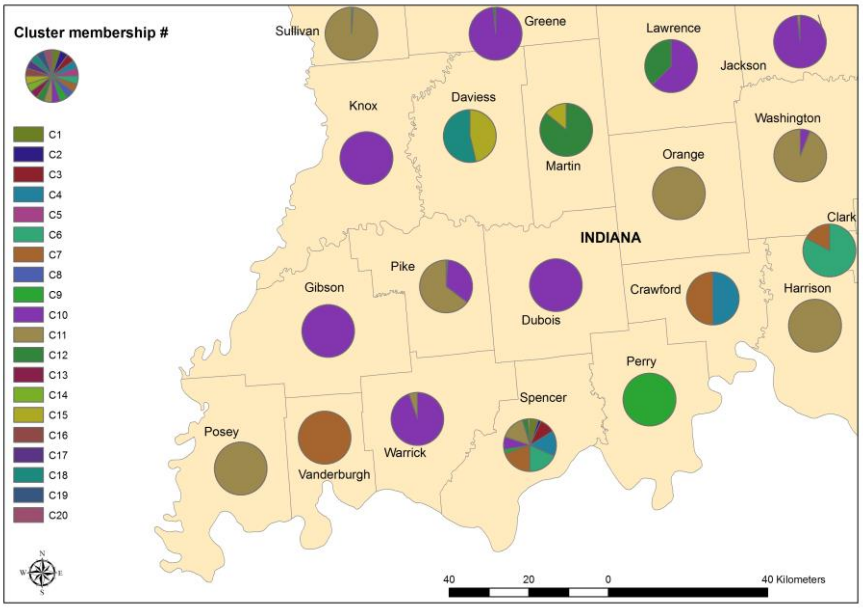


Fig. 2. Example of election cluster memberships for some states in Indiana.

The spatial element, namely location of counties within the various time zones, was expected to be an additional challenge to explore. This specifically included determining whether voting results from counties on the east coast, which would supposedly close their polls and announce their results first, would sufficiently predict the patterns in the other time zones where polls were still open. The question of whether the time zone differences could be used was the main issue investigated during the testing phase.

2.3 Extrapolating from known results

On the election night, the previous election's voting results are not used in the predictions - only the associated previous similarities between counties are retained, as captured by their memberships within clusters. Once the cluster memberships have been derived and data on registered voters has been obtained, the final step of producing live forecasts on election night involves using the clusters to extrapolate from a sample of known new election results, as the results come in. In the case of the US elections, the initial sample was expected to have a biased spatial component due to time zone differences rather than be distributed evenly across the country, as is typically the case in the South African elections.

For the US presidential elections, the forecasting algorithms had to be adapted to handle the electoral vote calculations instead of the proportional voting system used in South Africa. The forecasts are generated by the adapted algorithms by iterating through the following high-level steps, continuously updating as more results are known:

- Use the results from the counties declared, together with the number of registered voters and the cluster membership values, to predict the new voting pattern for the clusters and the new predicted turnout for each cluster.
- Use the predicted voting behaviour and predicted turnout for each cluster (calculated in 1), together with the number of registered voters and the cluster membership values, to predict the voting percentages for the undeclared counties.
- Combine the actual party votes from the counties declared with the predicted party votes in the undeclared counties, using only those counties within the state, to produce a forecast per party for the state.
- Using the forecasted party percentages in each state, compute the winner in each state and allocate the electoral votes from that state to the winning party (with slightly different algorithms for Maine and Nebraska which use congressional districts) .
- Aggregate the predicted electoral votes per party across all the states to obtain the predicted winner of the US presidential elections.

3. Data access and preparation challenges, testing and preliminary findings

Obtaining the GIS information on the US counties did not prove to be too difficult. County data was downloaded from the US Census Bureau's Cartographic Boundary Shapefiles - Counties for 2015. In addition, the historical voting results per county for both the 2008 and 2012 presidential elections were downloaded from the Data.Gov website. As indicated previously, the voter registration data was more difficult to obtain but this was sourced on a state by state basis using the available state websites. For states where no sources could be found, the registered voters were estimated from the turnout given in the historical voting results data listed above.

An initial test run was done using fuzzy clusters of US counties based on 2008 presidential election voting patterns, combined with simulated "time-stamped" (the order in which counties are declared) inflows of results from 2012 elections, and the clusters based on 2008 results appeared to predict 2012 outcomes well. Unfortunately, no actual "time-stamp" was available and hence an estimated "time-stamp" based on poll closure times in the various states together with some randomisation within time zones was used. The electoral vote results from this test run of the 2012 US presidential elections are given in Figure 3 and Figure 4 below.

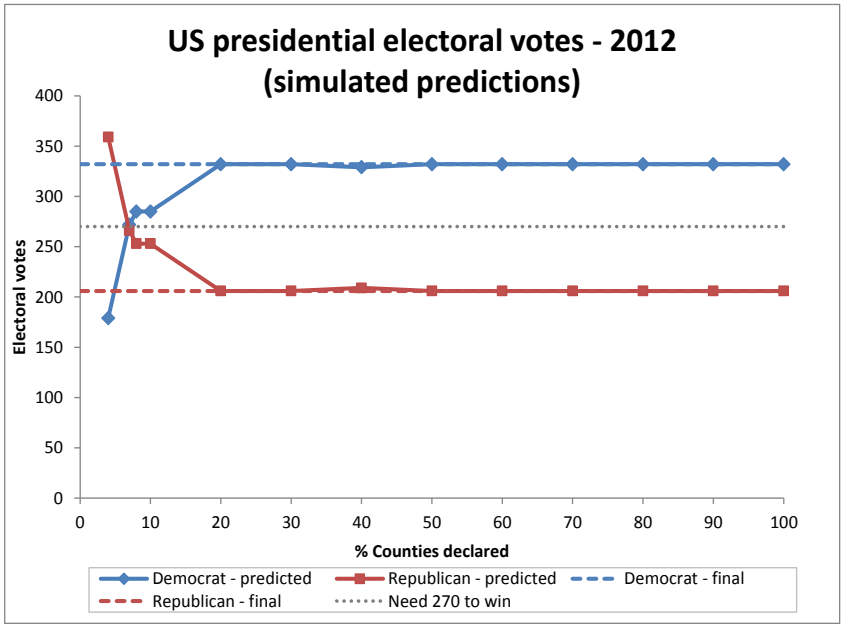


Fig. 3. 2012 US presidential electoral vote predictions at various percentages of counties declared.

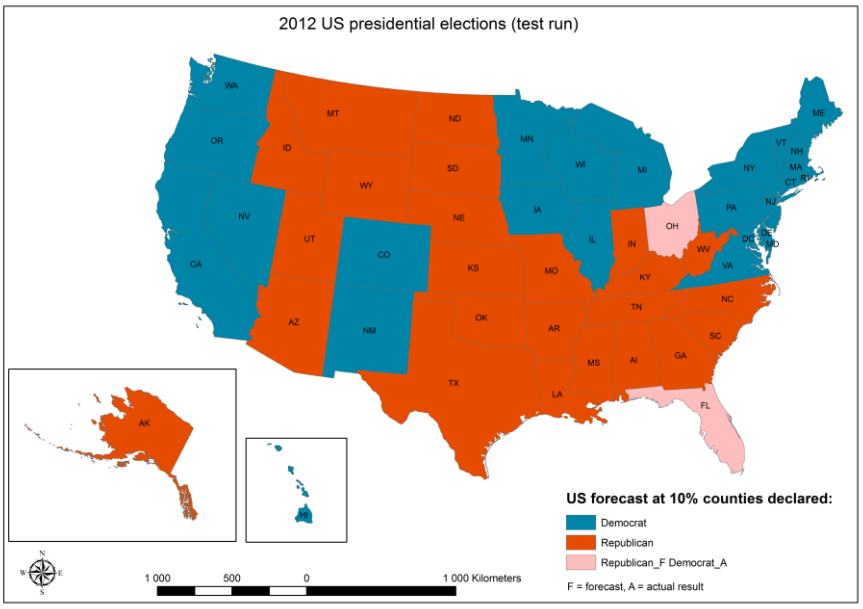


Fig. 4. 2012 US presidential election predictions per state at 10% counties declared nationally.

Based on the relative success of the simulated test results, it was decided that it would be possible to forecast the 2016 US presidential elections based on 2012 county voting patterns. Fuzzy clusters were therefore created using 2012 voting percentages per party in each county, as illustrated by the Indiana examples in Figures 1 and 2 and these cluster memberships formed the input for the 2016 predictions. Figure 5 shows an example of the spatial distributions of counties with relation to their main clusters. An example of a predominantly Republican, two predominantly Democrat, and a "mixed" (50-50) cluster are provided to show that these counties were spread across various US states.

The biggest challenge, however, was in obtaining the information on "live" actual votes, i.e. counted and declared votes per county. In South Africa, the Independent Electoral Commission collects all declared voting districts in one national database, but there appeared to be no such federal equivalent in the US. The team tried to obtain access to some of the US databases, but were not successful in this attempt. Eventually, the data released on the www.politico.com (Politico, 2016a) website was used for "near-real-time" data. Even though data on this website was released at county level soon after the county was declared, there was some time delay between the formal declaration and the data appearing on the website. The team was also not too sure about the format and the structure in which the voting results would be published on the website. It was found that a manual process of monitoring and updating results did not work well, and so a software tool ("data

scraper") was implemented to read data from the website and populate a database automatically at regular intervals. Prior information on data formats would have enabled the team to prepare the data scraping script before the election so as to enable rapid and continuous extraction of the results from the beginning until all counties were fully declared.

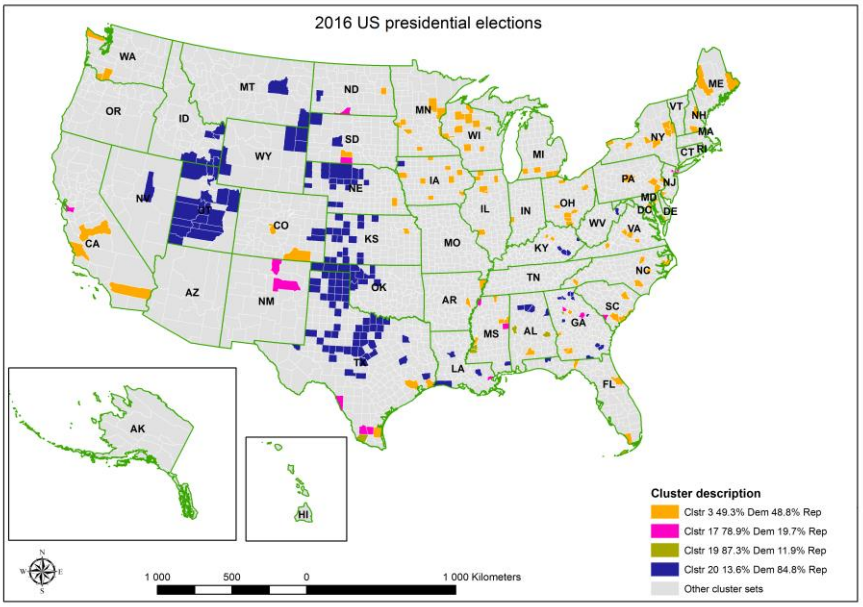


Fig. 5. Spatial distribution of counties with main memberships in one of four different clusters for use in the 2016 US presidential election predictions.

Early on the morning of the 9th November 2016 (South African time), the declared results were initially captured manually until the format and the structure of the results on the website could be incorporated into the data scraping script. The combined input from the manual capturing and the data scraper were used to make the first prediction after 8.7% of the counties were declared.

4. Final results and reflections

In general the predictions worked well since the model continuously predicted a win for Trump. However, the electoral vote predictions were not as stable as expected due to swings in the predictions of some of the closely contested states. The predicted electoral vote counts at various intervals of % counties declared are given in Figure 6, which illustrates some of those prediction swings. Figure 7 il-

lustrates the state predictions when 8.7% of the counties had been declared nationally and Figure 8 shows the distribution of these 8.7% counties declared.

It can be seen from the colour coding in Figure 8 of the 8.7% of counties whose results were available at that time point, that the sample was heavily biased in favour of Trump. However, although this sample bias did cause an over-estimation of the number of electoral votes that would go to Trump, the clusters in the model were still able to adjust for some of the bias and therefore still predicted 206 electoral votes for Clinton (26 votes under the final count). These counties were also clearly spatially biased, as expected, all being from the eastern half of the US. However, there were fewer counties declared from the far eastern side than expected. Despite this spatial bias, the majority of states throughout the US were predicted correctly, thus confirming that the model can still perform fairly well under such conditions. From Figure 7 one can see that there were only 6 states that were incorrectly predicted at this point.

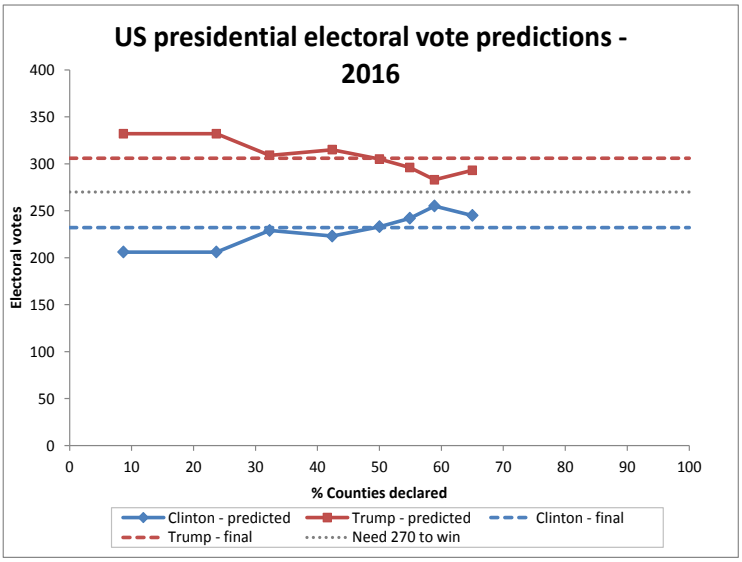


Fig. 6. 2016 US presidential electoral vote predictions at various percentages of counties declared.

Swing states according to Politico (2016b) are states that changed parties several times in the past five US presidential elections. The states as identified as swing states by Politico (2016b) are: Colorado, Florida, Iowa, Michigan, Nevada, New Hampshire, North Carolina, Ohio, Pennsylvania, Virginia and Wisconsin. The model had some difficulties in predicting the outcome in these states at 8.7% counties declared owing to the fluctuations between the Democrat and Republic wins of the counties in these states. The model managed to predict correctly at the 8.7% stage in the following swing states, namely Florida, Iowa, North Carolina, Ohio, Pennsylvania and Wisconsin.

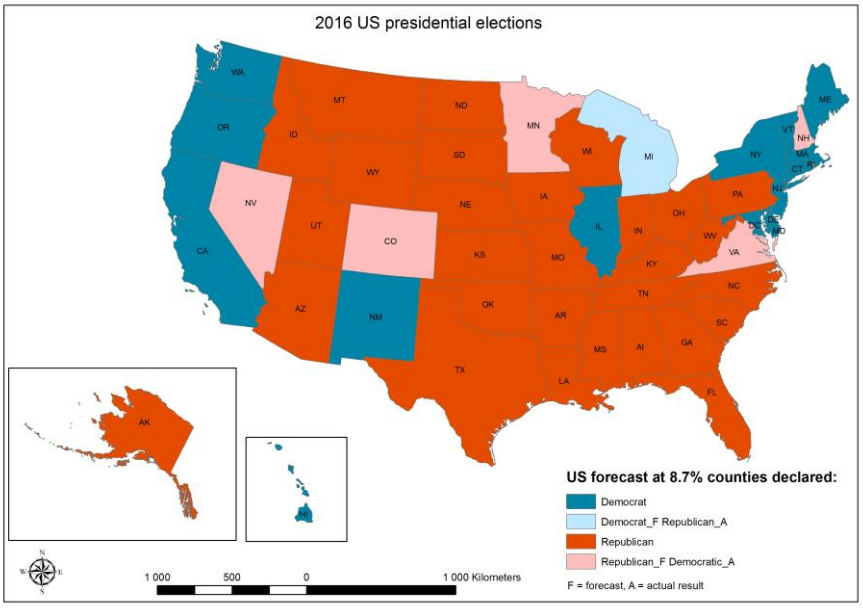


Fig. 7. 2016 US presidential election predictions per state at 8.7% counties declared nationally.

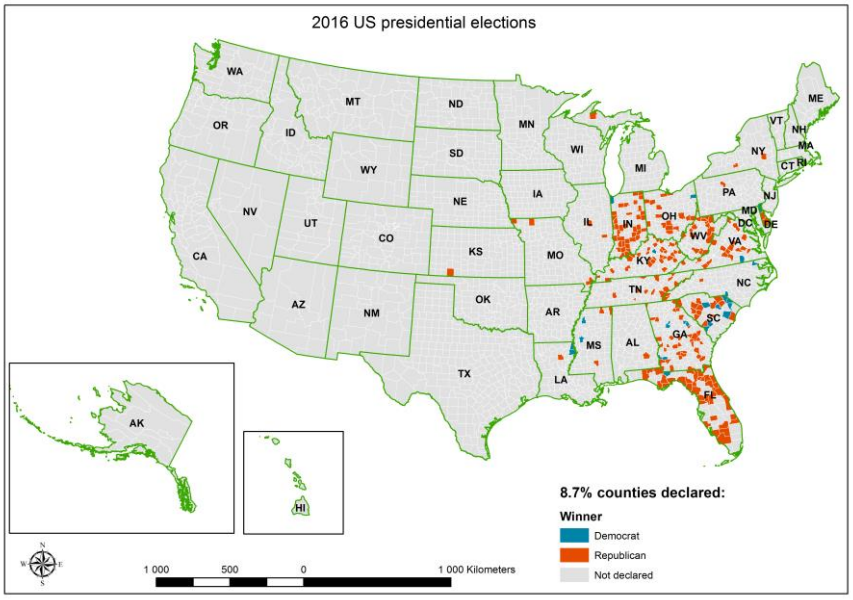


Fig. 8. Spatial distribution and political party preference of counties declared in 8.7% sample.

The swing states that the model did not initially predict correctly are Colorado, Michigan, Nevada, New Hampshire and Virginia. Minnesota, which was not listed by Politico as a swing state, was also not correctly predicted by the model at that point. However, the predictions were updated as more counties were declared, and predictions for the swing states improved. In general the model performed well at the 8.7% stage and was very close at roughly 30% declared. The overall outcome with Trump as a winner based on electoral votes was correctly predicted at the 8.7% stage and stayed close to the final count throughout updates, as illustrated in Figure 6.

In summary, the team was satisfied that the assumptions and methodology encapsulated in this election-night prediction model can be generalized and successfully applied outside of South Africa.

Acknowledgements

The authors would like to acknowledge Dr Jan Greben who developed the initial methodology and other members of the project team who worked through the night (South African time) on the US Election Day in order to help us test out the methodology on the 2016 US presidential elections, namely: Dr Zaid Kimmie, Paul Mokilane and Dr Ndumiso Cingo.

References

Greben, J.M., Elphinstone, C., Holloway, J., de Villiers, R., Ittmann, H. and Schmitz P.M.U, 2005, "Prediction of 2004 national elections in South Africa." South African Journal of Science, Vol 101, No. 3/4 p157-161, ISSN 0038-2353

Greben, J.M., Elphinstone, C. and Holloway, J., 2006. A model for election night forecasting applied to the 2004 South African elections. ORiON, 22(1), pp.89-103.

Politico, 2016a. 2016 Presidential Election Results . [ONLINE] Available at: http://www.politico.com/2016-election/results/map/president. [Accessed 8 November 2016].

Politico, 2016b. What are the swing states in 2016? Available at: http://www.politico.com/blogs/swing-states-2016-election/2016/06/what-are-the-swing-states-in-2016-list-224327. [Accessed 1 March 2017]