A Model for Election Night Forecasting applied to the 2004 South African Elections

Jan M Greben*, Chris Elphinstone and Jenny Holloway

Centre for Logistics, CSIR

Pretoria, South Africa

Abstract

A novel model has been developed to predict elections on the basis of early results. The electorate is clustered according to their behaviour in previous elections. Early results in the new elections can then be translated into voter behaviour per cluster and extrapolated over the whole electorate. This procedure is of particular value in the South African elections which tend to be highly biased, as early results do not give a proper representation of the overall electorate. In this paper we explain the methodology used to obtain the predictions. In particular, we look at the different clustering techniques that can be used, such as k-means, fuzzy clustering and k-means in combination with discriminant analysis. We assess the power of the different approaches by comparing their convergence towards the final results.

Keywords: Clustering, Forecasting, Elections

* *Corresponding Author:* Centre for Logistics, CSIR. PO Box 395, Pretoria 0001, Republic of South Africa, e-mail: jgreben@csir.co.za

1. Introduction

The South African elections present an ideal opportunity for analysts to carry out quantitative election night forecasts because of the excellent centralized and automated data collection during election night. Election results from the voting districts in which the counting process has been completed are immediately available at a central location, and the data available to forecasters are not limited to samples, as in some other countries (Morton, 1988; Karandikar et al. 2002). However, what makes these elections difficult to predict early is the fact that the early results are not representative of the final outcome because of the non-random order in which the incoming results are received. Therefore, there is a special need for developing methods that can counter this bias. Hence, the South African elections do not just demonstrate the need for forecasters, they also offer the forecaster the opportunity to test novel forecasting methods in a real-time application.

Various types of forecasts are carried out in countries engaged in democratic elections. In many countries the focus is on forecasts prior to the election. For example, in the United States websites proliferate before the presidential elections. Economic, social and political indicators are used to predict the outcome of the upcoming elections. For a survey of some of these analyses we refer to (Brown and Chappell, 1999). In the United Kingdom prior predictions have been based on economic and political factors (Lewis-Beck et al., 2004). That prior predictions can go seriously wrong was shown in the 1997 election in France (Jerome et al., 1999).

Another type of forecast, which is the topic of this paper, is the election night forecast. The relevance of such forecasts spans only a short period, namely between the closing of polls and the announcement of the final results. However, this is also a period of intense media interest, as the public is eagerly awaiting the results of the elections. Interviews with political leaders and panel discussions in the media add to this atmosphere of anticipation, and within this context rational, statistically based predictions can play a very useful role. In South Africa this atmosphere of anticipation is further enhanced by the strong bias in the early results, which is caused by the non-random order in which the results come in. This bias leads to a large variation in the actual percentage results with time. Hence, the public is eager to have access to more reliable predictions of the final

results. In view of this need for reliable election night forecasts, the South African Broadcasting Corporation (SABC), which is mainly responsible for the media coverage in election night in South Africa, sought the assistance of the CSIR to cover the 2004 elections. The CSIR had been involved in election night forecasting in 1999 and 2000 and the model that was used in the 2000 elections was again used to good effect in the most recent elections.

To determine which methods are most appropriate for election night forecasting in South Africa, we have to explain its electoral system in some more detail. Since 1994 South Africa has followed a system of proportional representation, in terms of which parties provide lists of candidates for the National Assembly and for each of the nine Provincial Assemblies. Seats are allocated from the top of each list, the number of seats gained by each party being proportional to the number of votes received by each party (Lemon, 2001). The elections are managed by the Independent Electoral Commission (IEC), which operates on election night from a central location in Pretoria. Since the number of seats is determined from the total number of votes across the country, the forecasters have to predict the final total number of votes for each party from the individual voting district results received up to that particular time. As explained below, the election-night forecasting model used for this is based on prior clustering of the voting districts. Although previous elections results are used to determine the clusters, they are not used as an input for - or an initial prediction of - the outcome of the current election. Notice that there are no exit polls in South Africa, on which early predictions can be based. In the United Kingdom the demands on the forecasters are slightly different. In that country a constituency system is used. Since most constituencies can be considered homogeneous in voter make-up, a large number of seats can be classed as “safe” and thus are unlikely to change unless a very large shift in voting allegiance takes place. Therefore, in that case one can use the results of previous elections as input into the analysis of the new elections, since the objective is to estimate the change in “share of vote” (Brown and Chappell, 1999). Other election night approaches, appropriate for their respective election systems, have also been developed for New Zealand (Morton, 1988), India (Karandikar et al., 2002) and Sweden (Thedeen, 1990).

As mentioned above, one of the main challenges to prediction of the elections in South Africa is that the early results are received in a very non-random way. For example, results from urban, more

affluent, areas tend to be available much earlier than those from rural areas. Since the voting behaviours in these different areas are also different, the early results are highly biased towards urban centres. Hence, the predicted final result cannot be based on simple projections from a small sample of early results as these early results are not representative. In other words: the usual statistical requirement of randomness, allowing an early call of the final result, does not apply. A successful prediction model will have to cope with this bias and we will try to demonstrate that a cluster model can be a very effective tool in this regard. In such a cluster model one divides the country into parts (clusters) with similar voting behaviour. As new results come in we can roll out the few votes counted in one cluster to the whole cluster, thereby getting a good estimate of the expected vote in this segment of the population. The first question we address in this paper is: how can we segment the electorate, given the available data and techniques? The second question addressed is: which cluster methodologies are suitable for such prediction models? There are many clustering techniques available in the literature. In our election night predictions we used the fuzzy c-means approach, advocated by Bezdek: (Bezdek et al., 1981; Bezdek, 1980; Nikhil and Bezdek, 1995). In the current post-analysis we have also analysed other clustering techniques. We assess the performance of different techniques by comparing their convergence to the final result.

The outline of the paper is as follows. In Section 2 we review the clustering methodology using the elections of 1999. In Section 3 we discuss the prediction formulae that can be used to predict the final outcome on the basis of early results. In Section 4 we discuss the convergence of the predictions to the final result for a number of different cluster technologies. Finally, in Section 5 we draw some conclusions.

2. Formulation of the Cluster Model

The purpose of the prediction model is to counter the bias resulting from the non-random order in which election results come in. To realize this objective we use a clustering approach. The cluster model aims to divide the population/electorate into groups with similar voting behaviour. The clusters are determined before the elections and are then used during the elections to extrapolate partial results to the whole cluster and thereby to the whole electorate. For the prediction model to converge as fast as possible towards the final results, it is essential that the electorate is clustered appropriately *before* the elections.

In order to construct the most appropriate clusters we have to consider two - partially related - questions. First we have to investigate which data are available on the electorate and decide which data can best be used in the cluster process. Second, we have to decide which cluster techniques are most suitable to construct an optimal prediction tool. Let us first consider the data question. In 1999 we had no suitable prior election available and we used demographic data to segment the electorate. At the time the most recent demographic and economic data were contained in the 1996 South African census. Since these data were available per voting district, they enabled us to design a cluster representation of the 1999 voting districts, which was subsequently successfully used for the 1999 elections. Because of the similarity of the 1999 and subsequent elections, we have been able to use the election results of the 1999 elections as a basis for the predictions in the subsequent elections in 2000 and 2004. The use of prior election data results in a more objective prediction tool than the earlier one based on demographic data, as the latter had to be supplemented by subjective assumptions on the importance of specific demographic and economic attributes on voter behaviour. We have also found that the predictions based on prior election data converge faster to the final results than those based on demographic data. Under certain circumstances it might be opportune to use a combination of the two data sources, however, we have not pursued this hybrid option so far.

In order to discuss the clustering methodology we need to establish some suitable mathematical terminology for the elections. In the national election of 1999 sixteen parties participated (more parties

participated in the provincial elections, but we will not consider these). Results are known for each voting district (see the Website (Website IEC)) and are indicated by:

$$x_{vp}, \quad p = 1,\ldots P, \quad v = 1,\ldots V. \tag{2.1}$$

Here p is the party index, while v represents the voting district. The total number of parties P increased from 16 in the 1999 election to 21 in the 2004 election. The total number of voting districts V was close to 15,000 in the 1999 elections, while in 2000 and 2004 it equalled 15,002 and 16,966, respectively. The results x_{vp} are expressed as percentages, and satisfy:

$$\sum_{p=1}^{P} x_{vp} = 100, \quad v = 1,\ldots V\ . \tag{2.2}$$

In addition to these results we know the number of registered voters N_v and the actual votes $N_v^{(a)}$ cast in each voting district (spoiled votes are not included in $N_v^{(a)}$). This information can be used to define the turn out:

$$T_v = N_v^{(a)} / N_v, \quad v = 1,\ldots V\ . \tag{2.3}$$

In order to construct clusters of voting districts with similar voting behaviour, we need to define the distance between the points x_{vp} in P - dimensional party space. We use the Euclidean measure:

$$d_{v_1 v_2} = \left\| \vec{x}_{v_1} - \vec{x}_{v_2} \right\| = \sqrt{\sum_{p=1}^{P} (x_{v_1 p} - x_{v_2 p})^2}\ . \tag{2.4}$$

This distance measure emphasizes bigger parties. If we want to emphasize smaller parties we could replace it by a standardized distance:

$$d_{v_1 v_2} = \sqrt{\sum_{p=1}^{P} \left(\frac{x_{v_1 p} - x_{v_2 p}}{\Delta x_p} \right)^2}\ , \tag{2.5}$$

where Δx_p is the standard deviation for party p, however, we have not used this measure in the current study.

The second question to consider is the choice of a suitable clustering technology to be used in the prior segmentation of the electorate. So far we have used the fuzzy clustering approach advocated by Bezdek (Bezdek, 1980), however, part of the current study is meant to compare it to other cluster

methods. In the fuzzy approach a suitable objective function is minimized, thereby optimizing the positions of the cluster centres so that the sum of distances squared between the cluster centres and the cluster members is minimal. This philosophy is similar to that in the k-means method [Kaufman and Rousseeuw (1990)], however, in the fuzzy case each element has a distributional, rather than a discrete, membership of the clusters. This distributional membership has distinct advantages in the present context, as it allows us to make predictions for *all* clusters, as soon as the first result is available. Also, the use of an optimization principle in the fuzzy method results in certain convenient properties of the mathematical expressions for the forecasts. The popular k-means method, on the other hand, is not based on a powerful optimization principle, but is easier to apply and interpret, as the memberships are either 0 or 1. In this paper we will consider both the application of the fuzzy and k-means method, as well as that of a hybrid method, which will be introduced at the end of this section.

Since the reader may be unfamiliar with the fuzzy cluster approach, and as we introduce a slight generalization of Bezdek's method, we review a few pertinent formulae. The idea is to minimize the objective function:

$$J_m(u,v) = \sum_{v=1}^{V} N_v^{(a)} \sum_{c=1}^{C} (u_{cv})^m (d_{cv})^2, \qquad m > 1, \tag{2.7}$$

where d_{cv} is the distance between the element x_{vp} and the (unknown) cluster centre v_{cp}. The memberships u_{cv} are distributional and add up to one:

$$\sum_{c=1}^{C} u_{cv} = 1, \qquad v = 1,\cdots V. \tag{2.8}$$

Our generalisation of Bezdek's method consists of the inclusion of the weight $N_v^{(a)}$ in the objective function. The objective function is minimized with respect to the cluster centres v_{cp} and the membership values u_{cv}. The resulting memberships and cluster centres can be expressed as follows:

$$u_{cv} = \frac{1}{d_{cv}^{2/(m-1)}} \Bigg/ \sum_{c'=1}^{C} \frac{1}{d_{c'v}^{2/(m-1)}}, \quad c = 1\cdots C, \quad v = 1,\cdots V \tag{2.9}$$

and

$$v_{cp} = \frac{\sum_{v=1}^{V} N_v^{(a)} u_{vc}^m x_{vp}}{\sum_{v=1}^{V} N_v^{(a)} u_{vc}^m}, \qquad c = 1 \cdots C, \qquad p = 1, \cdots P. \tag{2.10}$$

Since these expressions are mutually dependent, the set (2.9-10) is not a closed solution. As a consequence, we have to start the solution with an initial guess for the memberships u_{cv} or for the cluster centres v_{cp}, and iterate between (2.9) and (2.10) until we have reached convergence. No guarantee of obtaining a global solution can be given, in general (the situation is similar in the k-means case). The cluster centre v_{cp} has a natural interpretation: it is the average voting pattern in cluster c.

The role of the parameter m may require some elucidation. Different values of m refer to different ways of clustering the data. Obviously m has to be larger than unity (for $m < 1$ the optimization would maximize, rather than minimize, the objective function). In the singular limit $m \downarrow 1$ we recover the k-means case, where the memberships are either zero or one. With increasing m the clusters become fuzzier. For the extreme case of m being infinite, all elements have equal membership in each cluster; i.e. all clusters are identical. Hence, m characterizes the crispness of the solution. In the construction of our clusters we employed a value m = 1.2. In recent work we have tried to establish an optimal value for m by minimizing the difference between predicted and actual values of the voting district results in the 2004 elections. This has led us to a preferred value of 1.4. Bezdek has used the value 2 in some of his work (Nikhil and Bezdek, 1995). Notice that the method itself does not fix the value of m, as the objective function is optimized for any given m-value. The fuzziness or crispness of the cluster representation can also be captured by the so-called Dunn parameter (Dunn, 1976), defined according to:

$$F_c = \frac{1}{V} \sum_{c=1}^{C} \sum_{v=1}^{V} u_{cv}^2, \tag{2.11}$$

or generalized in the usual way:

$$F_c = \sum_{c=1}^{C} \sum_{v=1}^{V} N_v^{(a)} u_{cv}^2 \Big/ \sum_{v=1}^{V} N_v^{(a)}. \tag{2.12}$$

This expression has a maximum value of 1 for $m \downarrow 1$, i.e. for the k-means approach. For $m \to \infty$ the minimum value of $1/C$ is reached. One often defines the normalized Dunn number:

$$F_c' = \frac{C\,F_c - 1}{C - 1}\,, \tag{2.13}$$

that varies in a fixed range [0,1]. The 20 clusters for the 1999 elections have a normalized Dunn number of 0.85. Recently designed cluster representations based on the 2004 elections feature a normalized Dunn number of 0.76 for 40 clusters and 0.79 for 20 clusters.

In addition to fuzzy and k-means clusters, we will use clusters based on a hybrid approach, which combines the k-means procedure with discriminant analysis. The discriminant analysis serves two purposes. Firstly, it provides a criterion for selecting the best k-means clusters by comparing the error counts obtained. Secondly, the posterior probabilities for each element belonging to the different clusters can be used as a new definition of fuzzy memberships. In the discriminant analysis we used a parametric approach based on a multivariate normal distribution within each cluster/class. This allowed us to derive a linear discriminant function using the pooled covariance matrix (Seber, 1984; SAS/STAT User's Guide, 1990). One advantage of this hybrid approach is that it exploits the speed and simplicity of the k-means procedure in the determination of clusters and cluster centres. Another possible advantage is that the shared membership is easier to interpret than in the fuzzy and k-means method. For example, in the k-means method the memberships of an element lying nearly exactly between two clusters are still one and zero, while in discriminant analysis one would get values closer to 50%. For the fuzzy approach the situation is similar as in the k-means method if m is close to 1.

3. The calculation of predicted and expected results using prior clustering of the voting districts

In this section we show how we use prior clustering of the voting districts to assist us in the prediction of election results in a new election. Let us assume that at some point in time after the close of vote the first voting results come in. The set of voting districts for which results have come in at time t are

indicated by $\Omega(t)$. These 2004 results are indicated by y_{vp} to distinguish them from the 1999 results, which were indicated by x_{vp}:

$$y_{vp}, \quad p = 1,\cdots P_{new}; \quad v \in \Omega(t) \subset \Omega = \{ v = 1,\cdots V \} \quad . \tag{3.1}$$

The number of parties in the 2004 elections (P_{new} = 21) differs from that in the 1999 election (P = 16). There is no link assumed between prior and current parties, so the ordering of the parties is immaterial. However, the cluster index c does have the same meaning in the prior and new election.

In order to characterize the voting behaviour of cluster c we define a cluster centre in terms of the 2004 election results. It Is natural to use the following expression at time t:

$$v_p^{(c)}(t) = \frac{\sum_{v \in \Omega(t)} N_v^{(a)} u_{cv} \; y_{vp}}{\sum_{v \in \Omega(t)} N_v^{(a)} u_{cv}}, \quad p = 1,\cdots P_{new}, \quad c = 1,\cdots C, \tag{3.2}$$

in analogy to the expression for the cluster centre resulting from the minimization procedure, Eq.(2.10). Since, we are not bound by the expression u_{cv}^m in the current situation, we have used u_{cv}, as this leads to linear expressions in terms of the memberships, a distinct advantage, as we will see later. Equation (3.2) can easily be interpreted intuitively. The cluster centre for cluster c is an average of all available results y_{vp} at time t, weighted by the relevance (i.e. the membership and size) of each result with respect to cluster c. In the absence of typical cluster c results at time t, we still will be able to obtain a prediction for $v_p^{(c)}(t)$, as the finite memberships u_{cv} will link it to all available results y_{vp}. This is one of the advantages of the fuzzy clustering over k-means.

In order to distinguish these real time estimates of the cluster averages for the 2004 elections from the prior results from the 1999 elections, we have used a different notation for the cluster centres, namely $v_p^{(c)}(t)$ instead of v_{cp}. The only inputs taken from the prior clustering process in Eq.(3.2) are the membership values u_{cv}. Although the old cluster centres v_{cp} are not used in Eq.(3.2), they still play a role in characterizing the nature of the clusters. This characterization, for example in demographic terms, is useful when explaining the significance of the new cluster results to political analysts.

The effective turn-out in cluster c is defined by:

$$T^{(c)}(t) = \frac{\sum_{v \in \Omega(t)} N_v^{(a)} u_{cv}}{\sum_{v \in \Omega(t)} N_v \, u_{cv}}, \qquad c = 1, \cdots C \, . \tag{3.3}$$

By taking the average over the cluster results, weighted by the significance of each cluster to the uncounted voting district, we arrive at the following expression for the predicted result:

$$\hat{y}_{vp}(t) = \frac{\sum_{c=1}^{C} u_{cv} \; v_p^{(c)}(t) \, T^{(c)}(t)}{\sum_{c=1}^{C} u_{cv} \; T^{(c)}(t)} \qquad p = 1, .. P_{new} \, , \quad v \notin \Omega(t) \, . \tag{3.4}$$

The turn-out $T^{(c)}(t)$ is included in this expression to guarantee certain convenient properties of the aggregated results (see later). In the spirit of the fuzzy clustering expression we could have used u_{cv}^m, rather than u_{cv}, however, post election analyses have shown that u_{cv} gives better predictions than u_{cv}^m. In the definition of the cluster result, Eq.(3.2), we have also used the linear form.

The predicted turn-out for district v can be defined in a similar way:

$$\hat{T}_v(t) = \sum_{c=1}^{C} u_{cv} \; T^{(c)}(t), \qquad v \notin \Omega(t) \, . \tag{3.5}$$

Observe that all predicted values are supplied with a hat. The expressions (3.4) and (3.5), together with the known results over $\Omega(t)$, can now be aggregated over the whole country, or over smaller areas, like a province, metro or municipality. For example, for the whole nation we get the following prediction:

$$\hat{y}_p(t) = \frac{\sum_{v \in \Omega(t)} N_v^{(a)} \, y_{vp} + \sum_{v \notin \Omega(t)} N_v \hat{T}_v(t) \, \hat{y}_{vp}(t)}{\sum_{v \in \Omega(t)} N_v^{(a)} + \sum_{v \notin \Omega(t)} N_v \hat{T}_v(t)} \, . \tag{3.6}$$

We notice in passing that all predictions automatically satisfy the sum rule, Eq.(2.2), i.e. the total percentage of votes always equals 100%.

In addition to the *predicted* value, Eq.(3.6), one can also define the *expected* value at time t:

$$y_p^{exp}(t) = \frac{\sum_{v \in \Omega} N_v \hat{T}_v(t)\, \hat{y}_{vp}(t)}{\sum_{v \in \Omega} N_v \hat{T}_v(t)} \qquad p = 1,..P\,. \tag{3.7}$$

We can also calculate the expected value for a known voting district by applying Eq.(3.4) to $v \in \Omega(t)$. By comparing this expected value to the actual value one can assess the unexpectedness of the result in the voting district v. This could be useful for identifying possible fraud in the elections, or to identify results that are of special interest because of their extreme nature (outliers).

Let us conclude this discussion of the prediction formulae with a motivation for the inclusion of the turn-out coefficients in Eq.(3.4). If we calculate the expected value $y_p^{exp}(t)$ at the end of the voting process (i.e. when $\Omega(t) = \Omega$ at $t = t_f$), we obtain the non-trivial identity:

$$y_p^{exp}(t_f) = y_p^{act}(t_f)\,, \tag{3.8}$$

where the actual national result at time t is given by:

$$y_p^{act}(t) = \frac{\sum_{v \in \Omega(t)} N_v^{(a)}\, y_{vp}}{\sum_{v \in \Omega(t)} N_v^{(a)}}\,, \qquad p = 1,..P\,. \tag{3.9}$$

The expected and predicted values are not equal prior to t_f. The desirable identity, Eq.(3.8), is only valid if we employ the linear expression, Eq. (3.2), for $v_p^{(c)}(t)$ and include $T^{(c)}(t)$ in Eq.(3.4). It is an example of an identity which is possible thanks to the elegant mathematical basis of the formulation.

While the prediction formulae are cast into the language of fuzzy clustering, they can - and will - also be used for the other cluster methods analyzed in the following: the k-means and the k-means combined with discriminant analysis estimates of the memberships.

So far we have not discussed the choice of the number of clusters, C. Since there are no strong theoretical reasons for choosing one value or another, we have to test the performance of different values in practice. This can be done by means of measures (norms) which are defined independently

of the value of C. Such measures will be defined in Section 4. In our application of the model to the 2004 elections we have used 20 clusters. Generally, the more clusters we have, the more accurately we can cover all possible voting patterns. However, this comes at a price, as an increase in the number of clusters leads to a reduction in the predictive power. This is illustrated by the extreme case that each voting district has its own cluster: in this case no unknown result can be predicted, as the link between the unknown result and known cluster predictions is non-existent. The other extreme is that all voting districts belong to one cluster: in this case the cluster result equals the actual result, so that the predictions are identical to the actual result, and no correction of the bias takes place. Hence, the choice of the number of clusters must be a compromise between the ability to discriminate different voting behaviours and the potential to make predictions at an early stage.

We have analyzed a range of C-values in a post-election analysis, where we tested the predictions on the same data (2004 elections) that were used to construct the model. We found an improvement in terms of the aforementioned measures when we went from 10 to 20, and eventually to 40 clusters. However, this improvement may be a consequence of the fact that the test and calibration data were the same. We have also tested the number of clusters by using old calibration data (1999 elections) with new results (2004 elections) using the k-means method. Here we found that a number of 16 clusters is optimal. In summary, the number of clusters does not seem to be so critical for the predictive power as long as it is in the range 10-40.

In the previous paragraphs we discussed the number of clusters in terms of the predictive power of the resulting cluster model. One might also look at the demographic nature of the resulting clusters, and use this to characterize the voting behaviour of certain demographic groups, as this is where the media and the public is interested in. This leads to another set of criteria to choose the number of clusters. It is easier to keep track of a small number of clusters and comment on their behaviour in the new elections. On the other hand a large number of clusters allows one to identify smaller groups with characteristic demographics, and comment on these. So again we have to find a compromise between the advantages of large and small cluster numbers, and a number of 20 clusters seems to be a happy medium from the current perspective, as well.

4. Real-time predictions based on various cluster methodologies

In the previous section we derived various formulae for the prediction of the final election outcome on the basis of early results. We can analyze the convergence of the different methods visually, by comparing different graphs. The simplest way to do this is by providing the results for the three methods (fuzzy c-means, k-means and k-means with discriminant analysis) as if they had been used in the prediction of different parties in the national elections. In Figure 1 we show these predictions, as well as the actual results against the percentage of votes counted for the African National Congress (ANC):

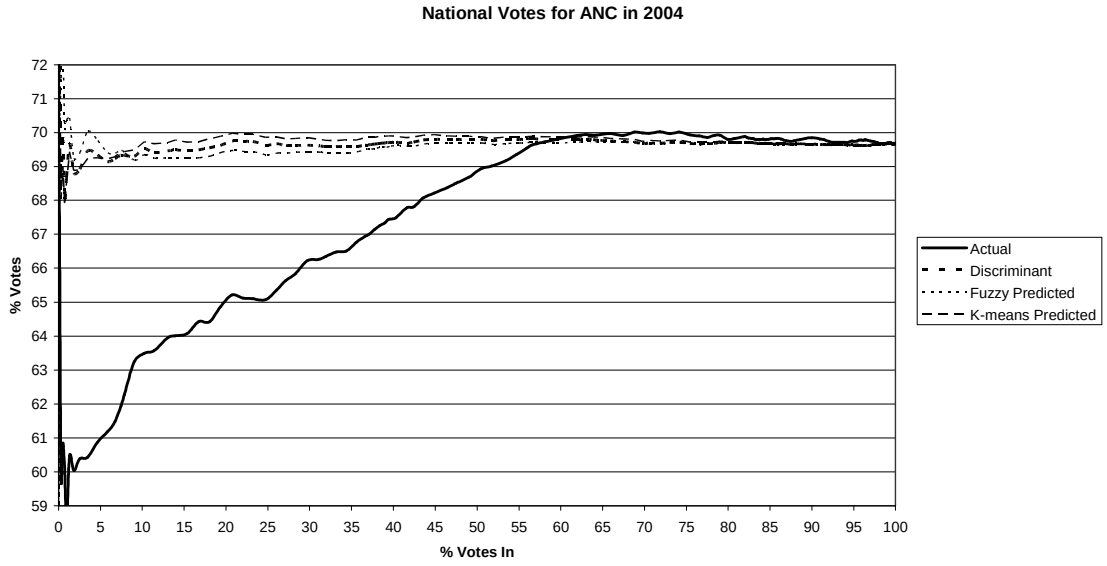

Figure 1 ANC results for the national elections and their predictions as a function of the number of votes counted

It can be seen that all the predictions have already converged to the final result when only a small percentage of the votes had been counted. At this stage the actual results are still far removed from the final results. In view of the more elaborate determination of the clusters in the fuzzy approach and the expected improvement by introducing the discriminant analysis over the k-means approach, we had expected there to be a gradual improvement by going from the k-means to the k-means + discriminant analysis, and finally to the fuzzy approach. However, in the case of the ANC results there is no clear evidence for this behaviour.

In Figure 2 we show the results for the second largest party, the Democratic Alliance (DA):

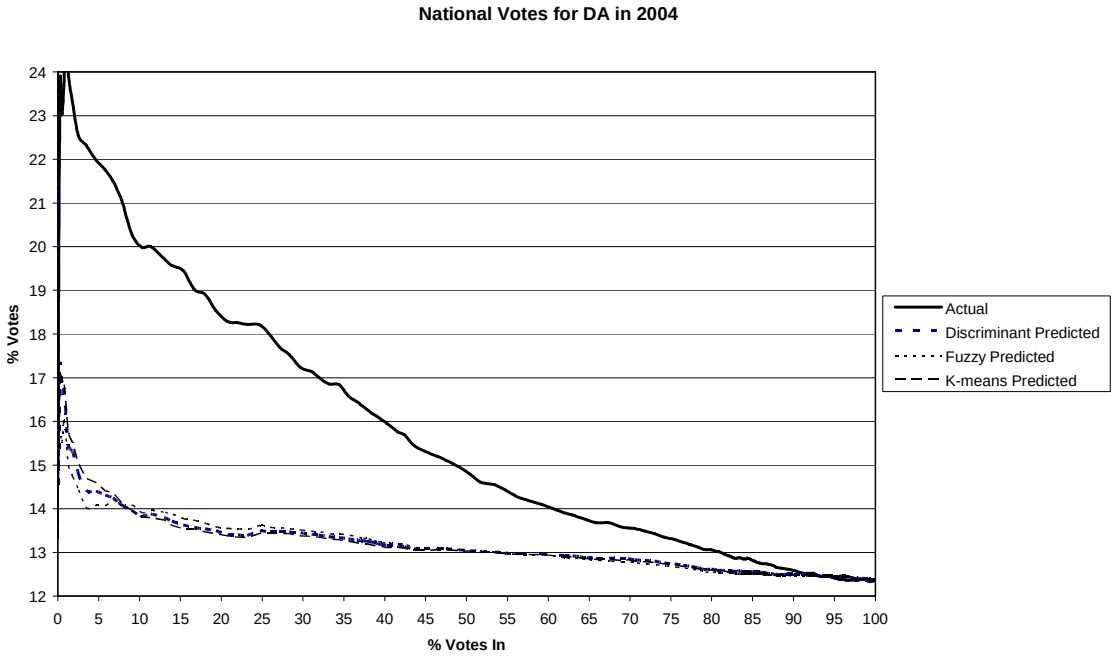

Figure 2 DA results for the national elections and their predictions as a function of the number of votes counted

Here it takes a little longer to produce a result close to the final one. However, again the different methods yield very similar convergence. From the start until about 7% of votes are in, the fuzzy calculation gives the best predictions. However, from 7% until 45% of votes in, the k-means prediction is slightly better. Beyond the half way point no discernible difference can be seen between the three predictions. Again the actual results converge much slower towards the final result.

Finally, in Figure 3 we show the results for the Inkatha Freedom Party (IFP). Here the fuzzy calculation is preferred throughout, the k-means predictions being the least effective of the three. This is the result that we had originally expected, as stated above.

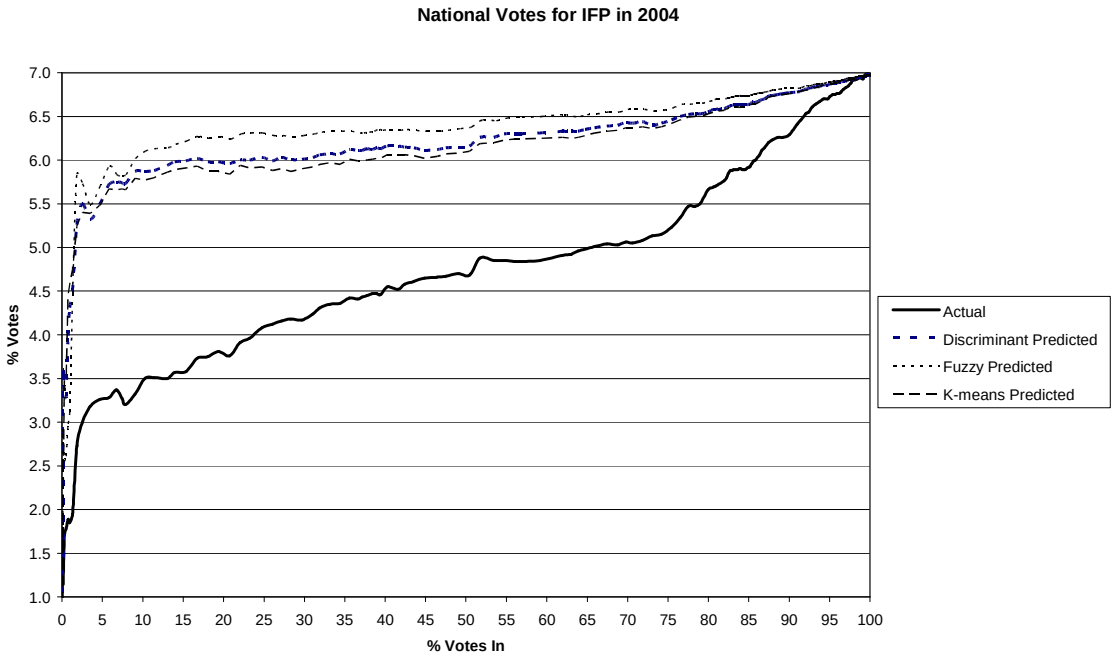

Figure 3 IFP results for the national elections and their predictions as a function of the number of votes counted

The examples of the three main parties in the elections illustrate the strong bias present in these elections. In the beginning the actual results give a strong showing for the DA and a weak showing for the IFP, if these are compared with the final results. The simple explanation for this phenomenon is that the DA voters are concentrated in the urban areas where votes are counted quickly, whereas the IFP supporters live mainly in rural areas, where votes are counted later. To some extent the latter explanation also shows the poor initial showing of the ANC. However, the effect is less pronounced here. The cluster prediction tools are clearly very effective in countering most of this bias.

Since, the individual party results are not completely decisive and consistent in deciding the effectiveness of the different approaches, as the relative differences between the three calculations are quite small, we have defined an overall error $E(t)$ to compare the three methods:

$$E(t) = \sqrt{\sum_{p=1}^{P_{new}} \{ \hat{y}_p(t) - y_p^{(final)} \}^2} \quad , \tag{4.1}$$

where

$$y_p^{(\,final\,)} = y_p^{act}(t_f)\,. \tag{4.2}$$

We can only calculate $E(t)$ after all results have come in, so it is only useful in a post-analysis. This error combines all 21 party results in the same way that we have constructed our clusters (namely using a Euclidean measure). Therefore, $E(t)$ is expected to display fewer fluctuations then the individual party results, and provide a more stable basis on which to judge the convergence properties of the three methods. The result is shown in Figure 4:

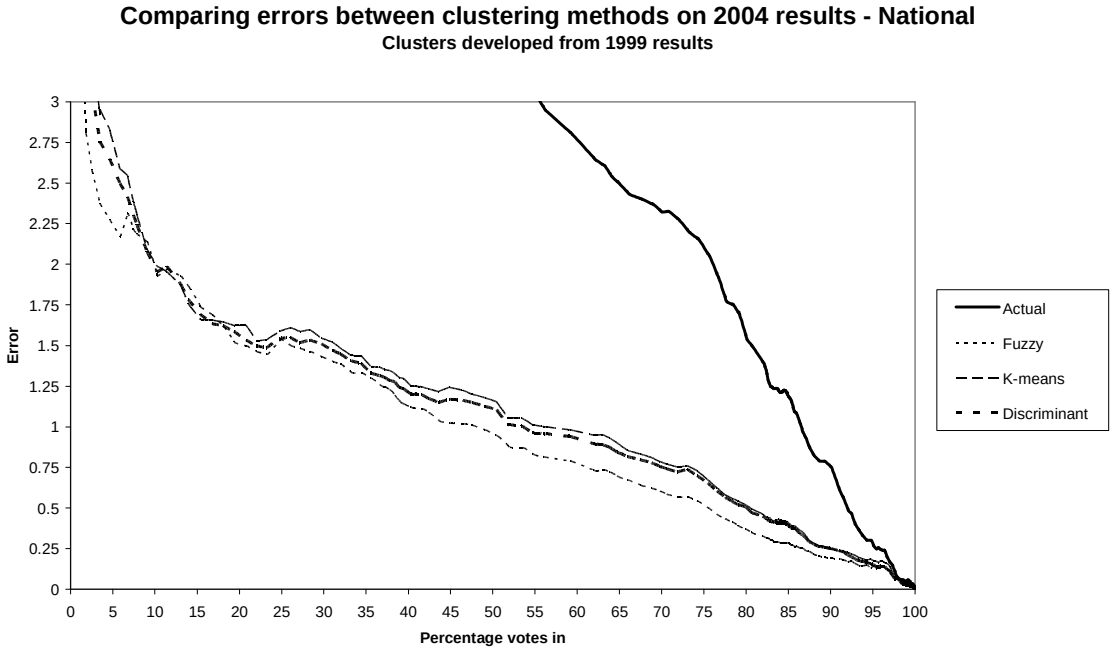

Figure 4 Comparison of average error $E(t)$ for three cluster methods used in the predictions

This graph displays the same tendencies as the IFP graph shown earlier: the fuzzy approach gives the best convergence and the k-means gives the worst convergence. The discriminant analysis shows some improvement over the k-means, but remains close to the k-means approach and is not able to bridge the gap between the fuzzy and k-means approach. However, by comparison with the actual results, all approaches seem about the same quality, and especially in the range of 5% to 20% of votes counted, there is hardly any difference.

Finally, we introduce a single measure that can be used to characterize the ability of the model to reproduce individual voting district results:

$$\chi(t) = \sqrt{\sum_{v\in\Omega} N_v^{(a)} \sum_{p=1}^{P_{new}} \{\hat{y}_{vp}(t) - y_{vp}\}^2 \Big/ \sum_{v\in\Omega} N_v^{(a)}}. \tag{4.3}$$

In contrast to the expression $E(t)$ in Eq.(4.1), $\chi(t)$ does not vanish for $t = t_f$. In fact, for $t = t_f$ this expression has special significance, since it represents the remaining difference between the expected and actual values when all results are known. Again, this quantity is only available in a post-analysis, as only the counted y_{vp} are available in real-time. $\chi(t)$ and $E(t)$ are good measures to compare methods employing different cluster numbers and methods, as they are not explicitly dependent on cluster numbers and system parameters such as m in the fuzzy approach. Below we give a table of the results for $\chi(t_f)$.

Table 1. Table of $\chi(t_f)$ values for various methods

Clustering used	$\chi(t_f)$
Fuzzy c-means (20 clusters)	14.92
k-means (15 clusters)	17.15
k-means + discriminant analysis (14 clusters)	16.88

It is clear that the fuzzy c-means scores best, whereas the k-means plus discriminant analysis does slightly better than the k-means on its own.

5. Discussion

The results in Section 4 indicate that a cluster model can be used to great effect for election night forecasting. However, the choice of the cluster method used to determine the clusters does not seem to play a major role. We compared three methods: the fuzzy c-means, the k-means, and the k-means

method combined with discriminant analysis. Two error measures were defined, which allowed us to compare the three methods in an objective way. The fuzzy method fared best under both measures. Taking the k-means error in Table 1 as a standard, we see that by adding the discriminant analysis component the error is reduced by 1.5% and that the fuzzy c-means method reduces the error by 13%. This confirms that the fuzzy c-means method is the best practical approach. However, since the differences between different cluster methods are so small, the choice of cluster technique remains mainly a choice of convenience and personal preference and familiarity. Our own preference goes out to the fuzzy c-means method, as it has a sound mathematical basis, contains the k-means approach as a special case, and also gives the best results, as we have seen.

Given the insensitivity to different cluster methods, one can ask whether there are other ways to improve the predictions. One possibility is to make better use of the counted election results in real-time. By using a dynamic clustering process, where one adjusts the clusters during election night, one might be able to use the real-time information more effectively. However, because of the real-time nature of election night forecasting, we need a robust method, so that it would be sensible to test such a delicate method first in a post-analysis. Another possibility is to use the prior election results as input into the current prediction process. At the moment this information is only used to construct the clusters. One could use prior elections result in one voting district as a – partial – guide for the behaviour of that voting district in the new election. By using trend matrices to link the old voting pattern to the new one, one can possibly improve the predictions. This possibility, which is less reliant on cluster techniques, is currently under study.

A final issue which can be raised relates to the confidence level of the forecasts. The issue, however, did not turn out too be of practical importance, since the forecasts are being updated so rapidly, that the degree of change is immediately obvious. Experience in the last three elections was that two features were required before confidence could be placed in the forecasts. These were that the variation should drop to the extent that the plot behaved smoothly with time, and that the graph does not display a constant increase or decrease. An example is the DA line in figure 2, which displays a negative slope, even after the prediction has turned smooth. An early claim on accuracy would then be unwarranted. The above argument is entirely intuitive and was applied via graphic inspection. The

authors have not as yet developed a more objective way of dealing with the issue, but this has not turned out to be in any significant way limiting the application.

A possible way of quantifying the confidence level at any point could be by measuring the deviation of the observed from the predicted results for the counted voting districts This could be done for individual parties and overall. The usual objection to such a procedure would be that the model is evaluated using the same voting districts as were used to calibrate the model, leading to an expected underestimation of the error variance. One response to this criticism would be to use a “hold out” sample. Since this would be computationally awkward in real-time, a more attractive solution would be to use the newly received voting districts before updating the model for validation. However, because of the bias it is not clear that the counted voting districts (even the most recent ones) could be considered representative of the areas which have not yet been counted and for which the predictions are being made. Further study of this issue is required to come to a solution that is both correct and practical.

References

BEZDEK JC, 1980, *A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms*. IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-2, no 1, 1-8.

BEZDEK JC, TRIVEDI M, EHRLICH R & FULL W, 1981, *Fuzzy Clustering: A New Approach for Geostatistical Analysis*, Int. J. Systems, Measurement and Decision, **1-2**, 13-24.

BROWN L. & CHAPPELL H, 1999, *Forecasting Presidential Elections using History and Polls*, Int. Journal of Forecasting **15**, 127-135

BROWN PJ, FIRTH D & PAYNE CD, 1999, *Forecasting the British Election night 1997.* J.R. Statist. Soc. A **162** Part 2, 211-226

DUNN JC, 1976, *Indices of partition fuzziness and the detection of clusters in large data sets*, in "Fuzzy Automata and Decision Processes", edited by M. Gupta, Elsevier, New York

JEROME B, JEROME V & LEWIS-BECK MS, 1999, *Polls Fail in France: forecasts of the 1997 legislative election*, International Journal of Forecasting **15**, 163-174

KARANDIKAR RL, PAYNE C & YADAV Y, 2002, *Predicting the 1998 Indian Parliamentary Election,* Electoral Studies **21**, 69-89

KAUFMAN L & ROUSSEEUW PJ, 1990, *Finding Groups in Data, an Introduction to Cluster Analysis*. New York, John Wiley & Sons, Inc.

LEMON A, 2001, *The General Election in South Africa,* June 1999. Electoral Studies **20**, 305-339.

LEWIS-BECK MS, NADEAU R & BELANGER E, 2004, *General Election Forecasts in the United Kingdom: a political economy model,* Electoral Studies **23**, 279-290

MORTON RH, 1988, *Election Night Forecasting in New Zealand.* Electoral Studies **7:3**, 269-277

NIKHIL P & BEZDEK JC, 1995, *On cluster validity for the fuzzy c-means model*, IEEE Trans. On Fuzzy Systems, **3**, 370-379

SAS/STAT User's Guide, Volume 1, Version 6, Fourth Edition, published by the SAS Institute Inc. (1990) Cary, USA

SEBER GAF, 1984, *Multivariate Observations*. New York, Wiley Series in Probability and Mathematical Statistics.

THEDEEN T. 1990, *Election Prognosis and Estimates of Voter Streams in Sweden*. New Zealand Statistician, **25**, 54-58

Website IEC, http://www.elections.org.za/Results/Elections99.asp, accessed 7 January 2005