Prediction of the 2004 national elections in South Africa

Jan M. Greben[a*], Chris Elphinstone[a], Jenny Holloway[a], Rosalie de Villiers[b], Hans Ittmann[a] and Peter Schmitz[a]

During the last three elections in South Africa, the CSIR was involved in the prediction of the final outcome on the basis of early results. In this paper, we describe the methods used by the CSIR in these elections and comment on the success of the model used. We compare the rate of convergence of our predictions towards the final results with the convergence of the actual results. We also comment on the special challenges and time pressures faced by a research team when it uses a scientific analysis tool in a real-time context. The performance of the system is determined by its ability to deliver accurate predictions at an early stage, since interest in predictions diminishes rapidly as the outcome becomes clear. In the event, our predictions proved to be very accurate and our forecasts played a vital role in the ability of the national broadcaster to 'call' the election in the hours after the voting stations closed.

Introduction

In 1999, the CSIR developed and implemented a model for predicting the final results of the 1999 South African national elections as results from the individual voting districts became available. In the following year, we applied a similar model to the 2000 municipal elections. Both projects were partially funded by the Independent Electoral Commission (IEC). For the 2004 elections, we were sponsored by the South African Broadcasting Corporation (SABC) and were allowed to use its facilities at the IEC headquarters in Pretoria. The model we used was essentially the same as that employed in 2000, but we updated it in terms of the 2004 voting districts. Since the SABC also requested predictions for the provincial elections in each province, we employed separate models for each of the nine provincial ballots. This enabled us to accommodate the different parties contesting the provincial elections and also allowed us to analyse the different voting behaviour patterns in each of the provincial polls. In this paper we describe the method used and illustrate the model with examples of the results obtained.

The elections in South Africa are rather special in that they are much harder to predict at an early stage than those held in most other countries. The normal situation is that the final result can be predicted accurately ('called') on the basis of actual results, once a sufficiently large percentage of votes has been counted. The percentages of votes that are needed can be based on simple statistical arguments. The problem with the early results in the South African elections is that they come in in a very biased way. For example, results in urban, more affluent, areas are recorded much sooner than those in rural, poorer areas. Since the voting behaviours in these different areas are also very different, the early results are highly biased. Hence, the actual results in the early counting stages are a poor indicator of the final outcome, so that one cannot use normal statistical arguments to estimate the closeness of the actual to the final results. A successful prediction model will have to cope with this bias.

A natural way of accounting for the bias in the order of incoming voting results is to divide the country in parts with similar voting behaviour. We can then roll out the few votes counted in one segment to the whole segment, thereby obtaining a good estimate of the total expected vote. This is the essence of the approach followed by us in all the elections. Following statistical jargon, we call these segments clusters. Clustering can be applied to a set of objects (in this case voting districts) which are characterized by certain attributes. These attributes can then be used to define distances between objects, allowing for a clustering of objects with similar attributes. The question is: which attributes are available in the case of voting districts, and which can best be used to characterize voting behaviour? In 1999, we did not have any usable results from the 1994 national elections, mainly because the voting arrangements in 1999 were very different from the earlier election. We therefore had to rely on demographic data on the voting districts (the 1996 census) to realize a segmentation of the electorate. We used a neutral policy towards this segmentation, assuming that all census parameters (race, language, income, education, rural vs urban, age, geographical coordinates) were equally important in the discrimination of different voting behaviours and thus should be weighed equally. Ideally, such assumptions should emerge from an investigation into the voting behaviour by explicit field work. However, there was neither time nor funding for such a difficult exercise. The model described above was used successfully in the 1999 elections.

Despite this success, we decided to follow another approach in the municipal elections of 2000. We now had access to a set of usable election results in the form of those from 1999. So, rather than using the 1996 demographic data as attributes with the uncertain demographic assumptions about their importance, we decided to employ the 1999 election results as attributes in our clustering procedure. Since a large number of parties (16) participated in the 1999 national poll, there was enough scope to distinguish between different voting behaviours. Clearly, clustering on the basis of voting behaviour would be more difficult in a two-party state, such as the United States, or a three-party one such as Great Britain, although it is certainly not impossible to use techniques similar to those discussed here. The use of election clusters rather than demographic clusters led to improved predictions (judged on the basis of their convergence to the final results) in the 2000 elections. In view of this success in the 2000 elections, we used the same election clusters in the 2004 elections. We did not update our clusters in terms of the 2000 election results, as the 2000 elections were municipal, whereas the 1999 and 2004 polls were national. It seemed to us that the advantage of using more recent results (by one out of five years) would not outweigh the disadvantage of basing national predictions on municipal elections. In principle we should also have used different cluster sets for each of the nine provincial elections but, because of the lack of historical data on these ballots and on account of severe time constraints, we did not do so.

The mathematical method used to develop such a system has to address various questions. First, we can choose between different clustering methods.[1] The main ones are the hierarchical, K-means[2] and fuzzy clustering (C-means) techniques.[3,4] We chose the last approach on account of its flexibility and superior mathematical properties. A numerical comparison between the two different methods is carried out in ref. 5. Another question is: what measure should be used to characterize the difference

[a]Centre for Logistics and Decision Support and [b]Information Society Technology Centre, CSIR, P.O. Box 395, Pretoria 0001, South Africa.
*Author for correspondence. E-mail: jgreben@csir.co.za

between different voting patterns? We chose to use the simple Euclidean distance between results, with each party contributing through the percentage of votes it received. Clearly, this puts more emphasis on the bigger parties, but these are the most important parties anyway.

The outline of the rest of this paper is as follows: first, we review the prediction model. Expressions for various quantities relevant in the elections are introduced and explained. Next, we assess the quality and usefulness of the model by comparing the predictions with the final results at different times after closure of the ballot boxes. We then make some observations on the challenges faced when a partially untested model is applied in a real-time media environment when speed is of the essence. Finally, we draw some conclusions.

Formulation of the cluster model

The purpose of the prediction model developed by the CSIR was to counter the bias in the order in which results came in. Hence, it was essential to form clusters in which the voting behaviours were similar, so that any available results within these clusters could be rolled out over the whole cluster. How can we ensure that the clusters represent similar voting patterns? One way is to assume that voting districts with similar demographics would vote in the same way. But, since there are many demographic attributes, the question is: which ones are mainly responsible for the voting patterns? Naturally, in view of South Africa's history, one might expect that race and language (in so far as they characterize certain cultural groups) would be dominant. However, even if such an assumption is correct, one would rather deduce it from the analysis of the results than assume it from the outset. A more objective approach is to look at the results of recent elections and link these to individual voting districts. On the assumption that these voting districts do not change too much from one election to the next, we can use the voting patterns in individual voting districts to characterize them. Since the voting districts have remained fairly similar since 1999, we have used this approach in our analysis of the 2000 and 2004 elections, by basing the clusters on the 1999 election results.

A popular cluster model is the so-called K-means method.[2] Here each object (i.e. voting district) belongs to one — and only one — cluster. A more natural approach is for each object to have a shared membership. This idea was introduced by Bezdek[3, 4] and called the fuzzy clustering approach, or C-means method. Mathematically, it is also better founded than the K-means approach, being based on the minimization of an objective function. Finally, the derivations of the different quantities relevant to the elections are easily expressed in the language of fuzzy clusters. Hence, we decided to use this approach in the development of the cluster model.

The objective function to be minimized is

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

The meaning of the different quantities is as follows. First, there are V voting districts (objects) labelled by the index v. The number of voters in a district is indicated by N_v. Each voting district is a partial member of each cluster c, indicated by the coefficient u_{cv}. These membership coefficients satisfy the sum rule:

$$\sum_{c=1}^{C} u_{cv} = 1, \quad v = 1,...V. \qquad (2)$$

The Euclidean distance between an object v and the cluster centre for cluster c is indicated by

$$d_{cv} = \left|\vec{x}_v - \vec{v}_c\right| = \sqrt{\sum_{p=1}^{P} (x_{vp} - v_{cp})^2}, \quad c = 1,...C \quad v = 1,...V, \qquad (3)$$

where x_{vp} refers to the percentage votes obtained by party p in the voting district v (P is the total number of parties in the 1999 elections). The cluster result v_{cp} characterizes the average voting pattern in cluster c, and is obtained by minimizing the objective function J_m. The other quantity fixed by the minimization procedure is the membership u_{cv}. Finally, m is a parameter that must be greater than 1 (typically around 1.4). It characterizes the crispness of the solution. In the singular limit $m \downarrow 1$ we recover the K-means result, i.e. objects are then only a member of one cluster. The fuzzy clustering process was carried out for the 1999 election results ($P = 16$) using 20 clusters ($C = 20$). In choosing 20 clusters we compromised between a large number (like 40) to allow for sufficient discrimination between different cluster centroids, and a small number (like 5) allowing predictions on a minimum of results.

The only information that we needed for carrying out calculations for the prediction system in the 2000 and 2004 elections are the memberships u_{cv}. The cluster result v_{cp} can be used to characterize the nature of the cluster and plays a role in the interpretation of the results and the identification of trends. Since the voting districts in 2000 and 2004 were slightly different from those in 1999, we calculated the memberships of new or modified voting districts in terms of the old memberships by using suitable geographical weighting procedures.

In order to characterize the nature of the clusters even further, we used the 1996 census data to obtain a demographic profile of the 20 clusters with similar voting patterns. In Fig. 1 we represent the first and largest cluster.

Let us now explain how the new election results were forecast once the first results had come in. We characterized the actual voting results in the 2004 elections by the same symbol x_{vp} that was used in the 1999 elections [see Equation (3)]. These results were sent to the IEC after the closing of the voting booths. The IEC then made them available to the SABC and to our system. The number of parties contesting the 2004 national elections was 21 (i.e. $P = 21$). The provincial elections had a varying number of parties (between 21 and 27). Because the SABC wanted the results per party (to be fed into its database), we treated all the parties separately, but reported on 12 parties explicitly.

Using the actual results up to time t, we constructed the expected percentages for the different clusters by means of:

$$v_p^{(c)}(t) = \frac{\sum_{v \in \Omega(t)} u_{cv} x_{vp} N_v^{(a)}}{\sum_{v \in \Omega(t)} u_{cv} N_v^{(a)}}, \quad p = 1,...P \quad c = 1,...C, \qquad (4)$$

where $N_v^{(a)}$ is the number of actual valid votes in district v. In order to distinguish these 2004 cluster predictions from the 1999 cluster centres v_{cp}, we used a slightly different notation. The total set of voting districts is indicated by Ω, whereas the voting districts which yielded results at time t form the subset $\Omega(t)$. As noted before, the membership values u_{cv} are those obtained from the clustering exercise in the 1999 election.

The effective turnout in cluster c can be defined in a similar way:

$$v_0^{(c)}(t) = \frac{\sum_{v \in \Omega(t)} u_{cv} N_v^{(a)}}{\sum_{v \in \Omega(t)} u_{cv} N_v}, \quad c = 1,...C, \qquad (5)$$

where N_v is the number of *registered* voters in district v.

We now predicted the unknown result in an uncounted district v by means of:

$$y_{vp}(t) = \frac{\sum_{c=1}^{C} u_{cv} v_p^{(c)}(t) v_0^c(t)}{\sum_{c=1}^{C} u_{cv} v_0^c(t)} \quad p = 1,...P, \quad v \notin \Omega(t) \text{ or } v \in \Omega - \Omega(t). \qquad (6)$$

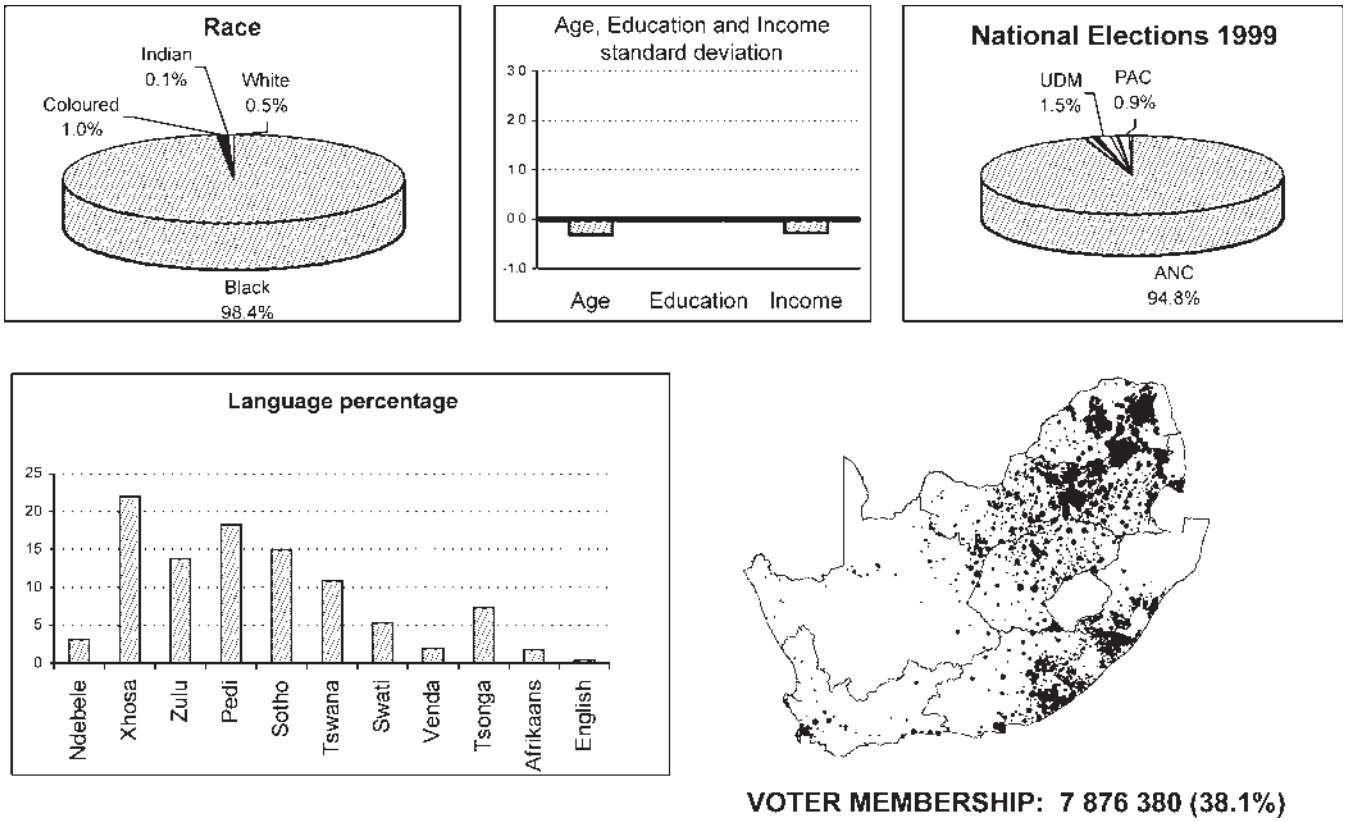

Fig. 1. Cluster 1: its composition is represented by the following elements: blacks (98.4%), urban (62%), age (33.1 years), annual income (R21 200), male (48%), African National Congress (94.8%).

The expected turnout in voting district v equals:

$$y_{v_0}(t) = \sum_{c=1}^{C} u_{cv}\, v_0^{(c)}(t) \quad v \notin \Omega(t)\,. \tag{7}$$

This result can also be used for the districts which had been counted, i.e. for $v \in \Omega(t)$. In this case it is called the expected, rather than the predicted, result. Large discrepancies between the expected and the actual results in individual voting districts may indicate either the inadequacy of our model or an unexpected or even questionable result.

The predictions for municipalities, metros, provinces and the country as a whole were then easily determined using suitable aggregations of known (x_{vp}) and unknown results ($y_{vp}(t)$) over subsets Ω_m of Ω. Notice that only the registered number of voters was known beforehand, so that the expected turnout must be used to guess the numbers of actual voters in voting districts in which the votes had not yet been counted.

Convergence of the predictions for the 2004 elections

In Figs 2–4 we display the comparisons between actual and predicted results for the national elections for the two main parties [the African National Congress (ANC) and the Democratic Alliance (DA)] and for one other party of special interest: the New National Party (NNP). The polls closed at 21:00 and counting of votes began immediately at individual voting stations. Results started to become available by 22:00. Ninety minutes after the first results came in (when 0.19% of votes had been counted), the predicted result for the ANC (Fig. 2) was 69.4%, nearly identical to the final result of 69.7%. Notice the big gap of 9% between the actual result and the final result in the early stages. This confirmed the bias resulting from the non-random order of the counting process mentioned earlier. It clearly shows the difficulty in using standard statistical arguments to predict the final outcome. Only after 60% of the votes had been counted (20 hours after the first results had come in) did the actual results coincide with the final result and the actual results could then have been used to 'call' the final result.

In Fig. 3 we show the results for the DA. These converged more slowly than the votes for the ANC. The results were accurate to within 1.5% when 4% of the votes had been counted (4 hours after the first results came in). It took another 4 hours (when 20% of the votes had been counted) before we obtained 1% accuracy. At that time the actual results still deviated by 6% from the final result. This convergence of our methods is similar to that in the 2000 elections, where we also reached 1% accuracy when 20% of votes had been counted. The slower convergence for the DA, compared to the ANC, may partly be because smaller parties are harder to predict than bigger ones. By redefining the distance measure in the cluster process to emphasize the results for smaller parties, one might be able to define other sets of clusters that would make it easier to predict the smaller parties. However, this may adversely affect the quality of the predictions for the larger parties. Another reason for the late decline in DA votes might be the late increase in the number of votes for the Inkatha Freedom Party (IFP), as the results from the rural areas of KwaZulu-Natal came in quite late.

Finally, in Fig. 4 we show the results for the NNP. These were of special interest because that party lost much of its support in the 2004 elections (dropping from 7.1% in the 1999 elections to 1.65% in the 2004 polls). Our original result of 1.6% (when approximately 5% of votes had been counted) compares remarkably well, since the final result (1.65%) differed only insignificantly from this initial result. On the other hand, the initial actual result (3.2%) was twice as large as the final result. This graph clearly shows that, even when support for a party changes dramatically, our prediction tool can deal with this change very well. The reason for this robustness is that our tool is based on the similarity of the voting behaviours of different members of the cluster. Hence, it does not matter whether the voters change their preferences radically between one election and the next. What *does* matter is that most members of the same cluster change in

the same way.

Similar results were obtained for the other parties. In particular the result for a new party, the Independent Democrats (ID), was accurate to within 0.2% when 5% of the votes had been counted. This again confirmed the robustness of our system.

To summarize the results: our prediction tool greatly speeded up the determination of the final voting results and most results were within 1% of the final result when 5% of votes had been counted (at 03:00). This was earlier than in the 2000 election, when such accuracy was reached only at 04:00. On the other hand, the DA results were predicted to within 1% at 04:00 in the 2000 election, whereas in 2004 we reached such accuracy only at 07:30. This was partly due to the slowness of the results coming in during the 2004 elections.

We have not shown the results for the provincial elections. The accuracy of these was often similar to those for the national election. However, in a few cases (Western and Northern Cape provinces) the results were not very satisfactory. This indicates that a reliable prediction for some of the provincial elections would require a separate set of clusters, specifically based on earlier election results for the province concerned.

Remarks on the operation of software in real-time media events

There is usually a great deal of media interest in elections. The 2004 election, marking 10 years of democracy in South Africa, was no exception. Public interest peaks when the polling booths have closed and the public is awaiting the outcome of the voting. This is the time when the public confronts its expectations (partly based on prior polling forecasts) with the actual results and learns who will govern the country for the next five years. Of additional interest are the fortunes of leaders and personalities of the smaller parties. This is also the time when the CSIR forecasts have to prove their worth.

The 2004 election marked the first time that the CSIR team had focused on the media, being sponsored by the SABC itself. Hence, there was a greater responsibility on the part of the CSIR group to deliver accurate, reliable predictions in real time. Traditionally, elections are a fertile arena for forecasts. The most common are the forecasts made prior to polling. These forecasts range from objective ones, such as opinion polls, to subjective ones based on expert interpretations and the speculations of political scientists. Naturally, these forecasts dominated the early discussions in the media on election night. However, as time passed, and the first election results came in, our predictions started to play an

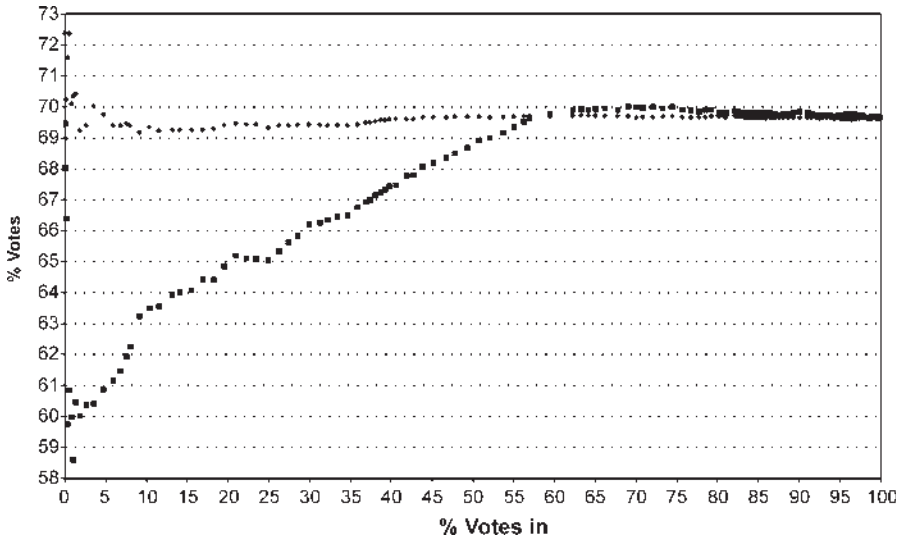

Fig. 2. Predicted and actual national results for the African National Congress as a function of the percentage of votes counted: ■, actual; ◆, predicted.

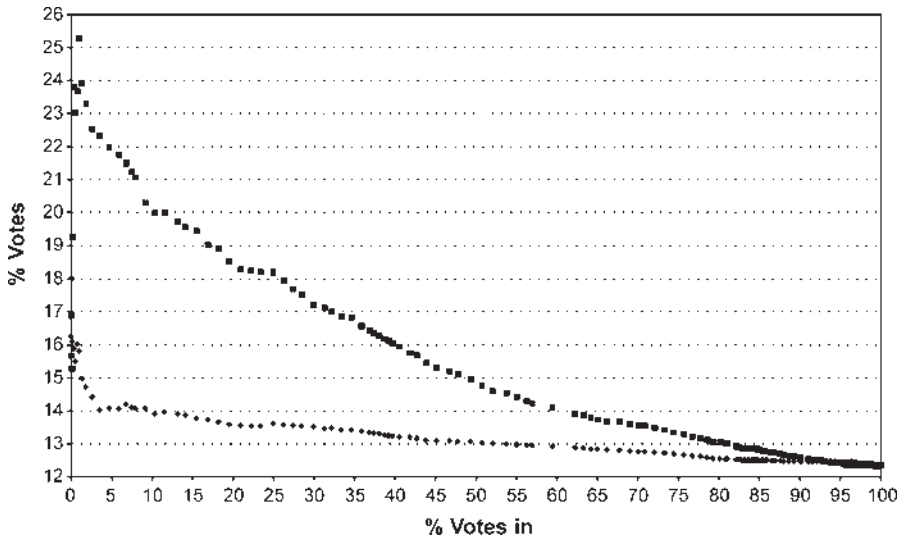

Fig. 3. Predicted and actual national results for the Democratic Alliance as a function of the percentage of votes counted: ■, actual; ◆, predicted.

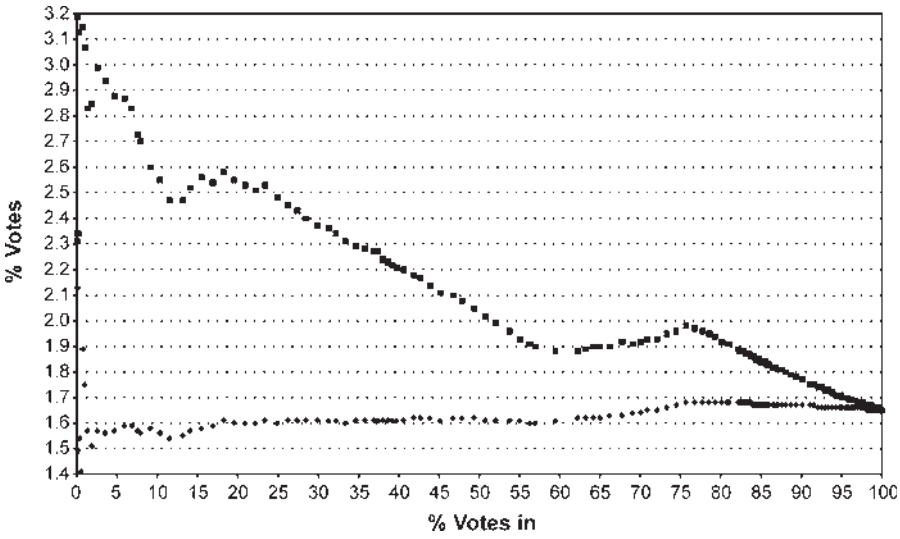

Fig. 4. Predicted and actual national results for the New National Party as a function of the percentage of votes counted: ■, actual; ◆, predicted.

increasingly important role in the discussions and media releases. Even the political analysts were eager to learn about our predictions in order to have a more scientific basis for their commentaries.

During election night in 2004, the SABC was the focal point for communicating the results to the country. It received the results from the IEC as soon as the latter had received them from the wards, and broadcast them via TV and radio. In addition, it hosted special all-night broadcasts that included analyses, interviews, panel discussions and opinions. The CSIR group was in constant dialogue with SABC analysts and provided the content of various media announcements. One item of much interest was the turnout, and in the early hours of the morning the SABC released the following media flash,[6] based on the CSIR's predictions:

> **Turnout expected to be 75%**
>
> *April 15, 2004 2:59 by Izak Minnaar*
>
> With just under 8% of the voting district results in, the CSIR predicts a national turnout of 75% in the elections.
>
> The provinces with the highest predicted turnout are Limpopo, Mpumalanga and Free State, all with 78%. The provinces with the lowest turnout are Western Cape with 72% and KwaZulu-Natal with 68%.

The prediction proved very accurate, since the final turnout was 75.5 %.

We also discussed our analyses, or summaries of them, in the television broadcasts and participated in panel discussions. Our spokesperson explained the rationale behind the model and the forecasts in these forums. It is essential to have a spokesperson on site for conveying the information in forms which could become part of the public debate. Clearly, the technical team had little time for this, as it had to address urgent issues concerning the operation and outputs of the system. One of the important objectives of the SABC was to 'call' the election as early as possible. To call an election is basically to announce the result before all votes had been counted. Obviously, this needs to be early enough to capture the interest of the public, but late enough to be accurate. In particular, as it relates to the smaller parties, it would be unacceptable to announce that a small party would win, say, one seat when ultimately it received none. In fact, even for the larger parties, the number of seats going to each party should not change between 'calling' the result and the final result. Clearly, our forecasts played a vital role in this calling process.

The SABC also had its own system that allowed various aggregations of the actual results. Since it had a link to a demographic database, it could also display the demographic characteristics of each voting district. Furthermore, the system could quantify the trends in the elections by having access to the 1999 and 2000 election results. The SABC displayed both the CSIR forecasts and its own results on the broadcaster's result system, and made its own independent analyses. Ultimately, it had to judge the consistency between — and the stability of — the different results and predictions appearing and decide whether it could 'call' a certain outcome.

Another role of the CSIR was to provide the SABC team with unexpected results. As stated before, these unexpected results can be identified by our system by noticing a particularly large deviation between the actual and the expected result. Some of the SABC's media flashes were based on these reports.

Summary and conclusions

The above results indicate that the actual voting pattern was initially strongly biased in favour of the DA and the NNP and against the ANC. This bias was also found in the predictions, although to a much lesser degree. In particular, in the case of the ANC and NNP predictions, the bias was largely removed by our system at the outset. In the case of the DA, the removal of the bias was less effective and, even close to the end of the voting process, a decrease in the DA's predicted results was continuing. Hence, although our model was successful in removing most of the bias, it was not completely immune to it.

We found that the quality of our predictions was not seriously affected by the antiquity of our clusters (based on the 1999 elections, five years prior to the latest poll), as the quality of our predictions was similar to that in 2000. In many cases the performance of the system was even superior to that in 2000. The quality of our predictions confirmed the conclusion reached after the 2000 predictions, namely that clustering based on previous election results provided a better basis for predictions than a demographic segmentation.

Since interest in predictions is short-lived, a question to ask is whether other aspects of our system can be exploited to supply useful information of more lasting value. One possibility is to identify trends in the elections. While the interest in the ballot is still high, it would be worthwhile if the system could provide real-time information on important trends in the voting. This was also an issue of great interest to the SABC. By visually inspecting the 1999 and (predicted) 2004 cluster results, one can form some qualitative assessments. However, such visual analyses are time-consuming, and difficult to carry out in real time. The use of a more automated system seems to be the solution. One problem is that the old and new results do not uniquely define the movement of voters from one party to another. However, by making suitable assumptions one can reduce these ambiguities, and come to a fairly automatic analysis of these trends. This aspect is presently under investigation[7] and has already led to interesting observations. An automated construction of the trends in the elections could possibly even be exploited in the prediction system. The trend analysis is also of potential interest to the political parties in order to judge where their support is coming from or, more particularly, where their support will be coming from in the future.

The use of such scientific methods in media events such as elections offers the scientist a rare, but important, opportunity to display science in action. The public is used to getting, and often ignoring, propaganda from the media in the form of advertisements and self-promoting spokespersons. The presentation of more scientific reports without the usual bias could give a refreshing angle on news events to which the public will pay attention. The responsibility of spokespersons is to present these technical methods and results in such a way that the public will appreciate the scientific work involved, without having to understand the methods in detail.

Received 3 December 2004. Accepted 21 February 2005.

1. Kaufman L. and Roussouw P.J. (1990). *Finding Groups in Data, an Introduction to Cluster Analysis.* Wiley-Interscience, New York.
2. MacQueen J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations, 5th Berkeley Symp. Math. Statist. Prob.,* eds L. Le Cam and J. Neyman, vol. 1, 281–297.
3. Bezdek J.C., Trivedi M., Ehrlich R. and Full W. (1981). Fuzzy clustering: a new approach for geostatistical analysis. *Int. J. Systems, Measurement and Decision* **1–2**, 13–24.
4. Bezdek J.C. (1980). A convergence theorem for the Fuzzy ISODATA clustering algorithms. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 1, 1 – 8.
5. Holloway J., Elphinstone C.D. and Greben J.M (2004). Forecasting the 2004 national election during the count: evaluating the model. In *Proc. 51st Annual Conference of the South African Statistical Association.*
6. South African Broadcasting Corporation (2004). Broadcast on SABC Radio by Izak Minnaar on 15 April 2004.
7. Greben J.M. (2004). Trend analysis applied to the 2004 national elections in South Africa. *Centre for Logistics and Decision Support Report* DP-2004/57, CSIR, Pretoria.