Measuring Transitions in Voting Pattern using Statistical Techniques

Working Paper
SHUBHAM SHUKLA, MANSI GOEL, SimplifyingStats.com

Statistical analysis of election results held within a short span of time helps in understanding the voting patterns and complex voting behavior of the electorate. The prevalent approach of using survey data for this analysis does not work very well where the diversity and density of the electorate could make this exercise costly and unreliable. Another possible approach could be to divide the analysis between geographical divisions and use the electoral results to understand the voting behavior. However such spatial analysis can only work if the geographical divisions are homogeneous which is not the case most of the times. To address such issues, we have developed a statistical framework using k Means Clustering algorithm and linear regression techniques. The methodology developed can be used both to do an ex post analysis of the result and as a tool to assist in essential groundwork before surveys are designed to predict election outcomes. We also extend our framework by suggesting potential interventions that could be designed using the model developed. The complete analysis has been carried out on data of Delhi Legislative elections held in December 2013 and Delhi Parliamentary elections held in May 2014.

1. INTRODUCTION

Political campaigns in India have dramatically changed with increase in use of technology in the electoral process which has now resulted in higher quality and quantity of voter data. National and regional parties have earlier relied mostly on extensive ground-work done by party workers to get voters point of view. Opinion polls and exit polls conducted by various survey agencies and predictions done thereafter have affected the nature of political campaign. Indian elections have always been considered extremely hard to predict [13]. Complex factors like block voting by caste and religion and lack of data-sets make predictions notoriously hard for experts in this field. Dramatic last minute swings have confound psephologist, for example, the complete failure of pollsters in predicting the 2004 parliamentary election. In fact none of the national election has been predicted within a satisfactory degree of confidence by any polling agency. There have been many reasons which have been attributed to the relative failure of pollsters in India but one of the main reason suggested is the difficulty in capturing representative data which can be used in statistic analysis. Multi-party democracy with first past the post system in a highly populated country requires sample size to be significantly higher than what can be afforded by the survey agencies. This work only uses the electoral data which is officially captured by the Election Commission of India and can be relied upon when compared to the survey data. Statistical analysis obtained from the work can be utilized in constructing better survey design which may lead to prediction of electoral results.

One of the key aspects of conducting election studies is to understand how voting pattern has changed over the period of time. Understanding the change in pattern helps the candidate in preparing an informed election strategy. Earlier similar insights were drawn from basic mean median analysis however now with huge amount of data available in public domain, new innovative methodologies can be developed to analyze what transpired during elections. Understanding the transition in voting pattern however becomes tough when there is no survey data available. Even if it is available, the authenticity of the survey methodology is questionable since often it is done by those organizations which are closely affiliated with parties themselves. Secondly, whenever elections are approaching or after election results are out, various organizations come up with their analysis with varied results based on survey data collected for the same area. Thirdly, unlike the developed countries where most of people are technology savvy and are quite

active on other social media platforms, there are places where demographics are such that major fraction of population is not active on social media. In such cases, online data collected does not truly represent behavior of people. Issues discussed above essentially create a problem of non-representative sampling of voters. When behavioral study and analysis is carried out on such a sample, results do not really help in understanding the voting pattern.

In this paper, we wish to come up with a framework that could be used to understand the transition in voting pattern keeping in mind the limitations of data discussed earlier. We introduce a model that relies on modern data mining techniques. We start with taking an example of Delhi legislative elections held in December 2013 and Delhi parliamentary elections held in May 2014 describing insights that could be drawn using traditional methods (**Section 2**). We then introduce our framework using various data visualization techniques while making following contributions in subsequent sections:

—We implement k-means clustering algorithm to capture similar behaving constituencies instead of depending on spatial analysis or geographical boundaries (**Section 5**).

—Groups obtained from cluster analysis are further fed into linear regression algorithm to come up with transition matrices (**Section 6**).

—To further support the utility of the model developed, we also describe how potential inferences could be drawn using this methodology and how better predictions of election results can be done (**Section 7**).

2. RELATED WORK

Few researchers have utilized the availability of in-depth demographic data and have used advanced statistical methods to understand electorate behavior. On one hand, [12] has used statistical methods to target voters, [2] on the other hand has used Twitter to monitor political sentiment and predict elections. Ref. [3] uses campaign finance data to measure ideology of the candidates. Ref. [16] has used supervised learning algorithms over readers' subjective reactions to news opinion articles to classify articles as liberal and conservative.

Researchers have also used Twitter data to predict election results. [14] has used streaming Twitter API to collect tweets and has proposed a prediction model to suggest the feasibility of replacing traditional polling methods. At the same time, scientists have also warned about the bias in the Facebook and Twitter data [1; 11].

Our study realizes the fact that Twitter data may not representative in nature. We in turn come up with a framework that results in matrix representing a transition. In social sciences, this is the canonical formulation of the ecological inference problem where one aims to extract information about individual behavior starting from information reported only at an aggregate level. Good overview of the approach considered for that problem can be found in [15].

3. UNDERSTANDING DELHI RESULTS: TRADITIONAL WAY

In this work, we explain the statistical framework by taking an example of elections held in Delhi within a span of 5 months approximately. In India, legislative elections are held for electing members of State Legislative Assembly while parliamentary elections are conducted across the nation for electing members of Parliament of India.

3.1 Political Scenario of Delhi

Delhi is the National Capital Region which is one of the states of India which also houses the central government. The region is divided into 7 parliamentary (central) constituencies (Fig 1), each of them have 10 assembly constituencies (state) dividing complete Delhi into 70 areas. The seven Parliamentary constituencies in Delhi are Chandani Chowk, East Delhi, New Delhi, West Delhi, North East Delhi, North West Delhi and South Delhi.

There are three major political parties in Delhi. Two of them are national level parties - Indian National Congress (INC) and Bhartiya Janta Party (BJP). INC positions itself as left of center party and has a strong base amongst the

Table I. Votes share in Assembly Elections(2013) and Parliamentary Elections(2014)

	2013	**2014**
INC	24.55%	15.10%
BJP	33.07%	46.10%
AAP	29.49%	32.90%
Others	12.89%	5.90%

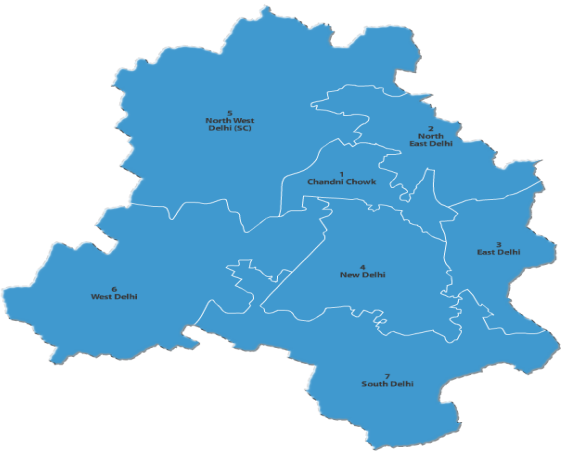

Fig. 1. Delhi Parliamentary Constituency Map

poor and minorities, had been ruling Delhi for 15 years till December 2013. BJP is perceived to be a right of center party which has traditionally done well in the urban areas. Third is a regional party which is a recently formed political entity - Aam Aadmi Party (AAP). This party positions itself on the plank of anti-corruption. Incumbent Government of Delhi led by INC had been facing public outrage because of various corruption cases against their leaders across India. Table I shows that AAP indeed made an impressive debut leading to fractured mandate in Delhi Legislative elections (2013). AAP formed Government in Delhi with the help of an unsolicited support of INC which lasted 49 days. Till the first election that is state assembly elections, AAP was competing with BJP for an urban vote base which was disenchanted with the ruling INC in state and center due to its alleged corruption. AAP was perceived to take leftward turn after coming to power and most of its election campaign targeted BJP rather than the Congress.

The results for Delhi in the Parliamentary elections (2014) were not so encouraging for AAP. BJP swept across all parliamentary constituencies as shown in Table II. Such a transition happened within a period of 5 months. Main thrust of this paper is to build a robust statistical framework which can be used to understand this process of transitioning. The methods developed in this analysis are generic and can be used in other elections as well.

3.2 Results

The results of the Assembly Election held in 2013 and Parliamentary Election held in 2014 indicate a complex voter behavior which have been tabulated in Table I and Table II.

Table I also shows that a budding regional party AAP gave a tough fight to BJP by securing 40% of the assembly segments while the latter gained majority winning 44% of all in Assembly elections 2013. However, in parliamentary elections, results declared for Delhi region on 16^{th} of May 2014 suggest that there was a huge voter swing. BJP won 85% of the assembly segments while AAP could obtain only 14% of assembly segments. INC could not create major

impact in either of the elections. Percentage of assembly segments won by BJP got almost doubled. On comparing the actual vote share (Table I), it is observed that vote share for both parties (AAP and BJP) increased. In legislative assembly polls (2013), Congress had secured a vote share of 24.55% which came down to 15.1% per cent in the Parliamentary elections. The BJP had secured 33.07% per cent vote share in the assembly polls which increased to 46.1% per cent in the Lok Sabha polls. The AAP's vote share also increased to 32.9% per cent from 29.49% per cent in the assembly polls.

Table II. Assembly Segments Won

	Delhi Assembly Polls 2013			**Lok Sabha Polls 2014**		
	BJP	**AAP**	**INC**	**BJP**	**AAP**	**INC**
Chandni Chowk *	3	4	2	9	1	0
East Delhi	3	5	2	9	1	0
New Delhi	3	7	0	10	0	0
North East	5	3	2	8	2	0
North West #	5	2	2	7	3	0
South Delhi	7	3	0	7	3	0
West Delhi +	5	4	0	10	0	0
Total	**31**	**28**	**8**	**60**	**10**	**0**

* A Janta Dal candidate won from Matia Mahal in the Assembly elections 2013
\# An independent won from Mundka in the Assembly elections
\+ A Shromani Akali Dal (Badal) candidate won in the Assembly elections 2013

3.3 Inferences

The traditional analysis which only focuses on macro level seat share and voting pattern can be used to draw inferences from the result. As suggested earlier, the voter percentage of AAP has increased in the parliamentary election as compared to assembly election. BJP increased its position both in terms of seats as well as vote share. INC performed poorly in both the elections. There are several possible ways for this to happen and only this information may not be sufficient to draw conclusions. For example, the increase in voter percentage of AAP suggests that the vote obtained in the 2013 election remained intact. The opinion on the ground suggests that the urban electorate who voted for AAP in 2013 moved to the BJP while AAP gained the traditional voters of Congress. The results in Table I and Table II are insufficient to conclude the same.

If the results are dissected to the geographical divisions and if the geographical divisions are somewhat homogeneous, further inferences can be drawn from the result. Out of the seven constituencies of Delhi, North East Delhi, North West Delhi and South Delhi generally comprise of many of the rural and sub urban areas. It can be seen from Table II that BJP strengthened its position in urban regions and hence improving from losing side in assembly elections to almost complete sweep in parliamentary elections. However, same sweeping results were not observed in rural regions although electorate change still favored BJP.

The analysis can be done at a more granular level which is at the polling booth level. In next section, we describe the data collected at the polling booth level and format of the same. We also explain the statistical test conducted over the granular data in our attempt to explain the transition in voting pattern.

4. COMPARING VOTING DISTRIBUTION: POLLING STATION LEVEL

Voting in Indian elections is done at polling station which is area-wise smallest unit for conducting voting in that area. The Election Commission of India publishes the electoral results at the polling station level which would be used in remainder of the analysis [5]. The advantage of using the polling station level data is that the voters within a single

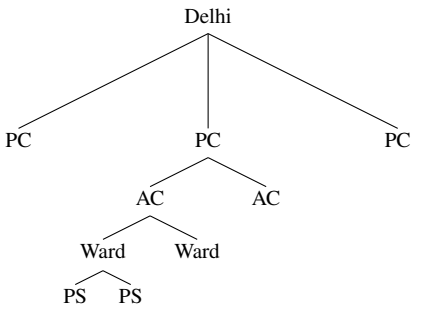

Fig. 2. Divisions for conducting elections

polling booth generally have similar economic, social and cultural background. Change in the voting behavior of a polling station is indicative of the general transition in voting pattern.

Illustrative tree structure below gives an idea about how geographical divisions are made so as to conduct election process in organized manner. Any **Parliamentary Constituency (PC)** has more than one **Assembly constituencies (AC)** in it. An assembly constituency is further divided into **Wards**. Each **Ward** has multiple **Polling stations (PS)** (See Fig 2).

4.1 Data Preparation

Raw data has been obtained from website of Delhi division of Election Commission of India. It has been cleaned and then consolidated in the form of two 2-D matrices, one for each election held on December 2013 and May 2014. Each raw data file contains a table which has as many columns as individuals contesting for that particular area. Snapshot of same is shown in Table III. For Delhi region, each 2-D matrix is transformed to 4 column matrix where 3 columns are for BJP, AAP and INC. Fourth column represents total number of valid votes in a particular area. Since data obtained is at the level of polling station, each row in matrix represents votes obtained by a candidate in that particular poll station. In total, we have obtained voting distribution for 10,335 polling stations.

4.2 Mathematical Notations

Suppose there are N_i^1 voting individuals in i^{th} polling station in the first election. Each individual has k_1 voting options to choose from. Election results for i^{th} polling station can be denoted by $y_i^1 = (y_{i1}^1, ..., y_{ik1}^1)$. Similarly, for the same polling station, results for second election can be denoted by $y_i^2 = (y_{i1}^2, ..., y_{ik2}^2)$ where k_2 is number of voting options available. These two equations can be combined to denote pair of elections as $y = (y^1, y^2)$ where $y^1 = (y_1^1, ..., y_n^1)$ and n represents number of polling stations. To understand how voting distribution for each party transformed in the elections, we have used Kolmogorov – Smirnov Test

Table III. Snapshot of Election Result Matrix

PS NO.	Candidate A	Candidate B	Candidate C	Candidate D	Candidate E	None of the above	Total valid vote	No. of rejected votes	Total No. of tendered votes
1	181	143	0	0	77	0	427	0	0
2	176	145	0	0	47	3	385	0	0
3	238	98	1	1	31	2	388	0	0
4	201	164	1	0	31	1	413	0	0

4.3 Kolmogorov Smirnov Test

There can be many simple analysis which can be done to understand the transition in voting pattern. One of the possible approaches is to compare the frequency distribution of the percentage votes polled in the polling booth for the three dominant parties.

In statistics, the Kolmogorov Smirnov test (KS test) is a non-parametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test). The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case). In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted [9; 10].

The two-sample KS test is one of the most useful and general non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. It may be used to test whether two underlying one-dimensional probability distributions differ or not [8]. In this case, the KolmogorovSmirnov statistic is as follows:

$$D_{n,n'} = sup|F_{1,n}(x) - F_{2,n'}(x)|$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively, and *sup* is the supremum function.

In our case, we have denoted pair of elections earlier as $y = (y^1, y^2)$ where $y^1 = (y^1_1, ..., y^1_n)$ and n represents number of polling stations. To apply KS test over pair of elections, we first summarize the results as $p^l_j = (p^l_{1\,j}, ..., p^l_{nj})$ for $l = 1,2$ and $j = 1,2\,and\,3$ i.e. BJP, AAP and INC where $p^l_{ij} = y^l_{ij}/N^l_i$ and i represents . KS test is then performed over voting distribution for each each party. For example, values represented by p^1_{i1} has BJP voting ratios across all polling stations in assembly elections 2013. This along with p^2_{i1} will be inputs in one of KS tests. Similar inputs can be designed for AAP and INC to understand how distributions for each party differ. Each KS test represents a change in voting pattern for each party.

Table IV below shows the result of KS tests conducted. Fig 3 shows how two voting distributions differ for each party. Maximum vertical deviation between the two curves is reported as D statistic in the KS-test. The results suggest that there is a significant difference in the distributions for all the three parties. The results also indicate that the distribution of AAP is fairly stable. This can be either because the voters of AAP remained intact with the party or there was an equal inflow and outflow in the voters which resulted in a similar distribution. This question would be analyzed in detail in the next two sections.

Table IV. KS Test Results

Party	**p–value less than**	**D–Statistic**
BJP	2.2e-16	0.382
AAP	2.2e-16	0.15
INC	2.2e-16	0.344

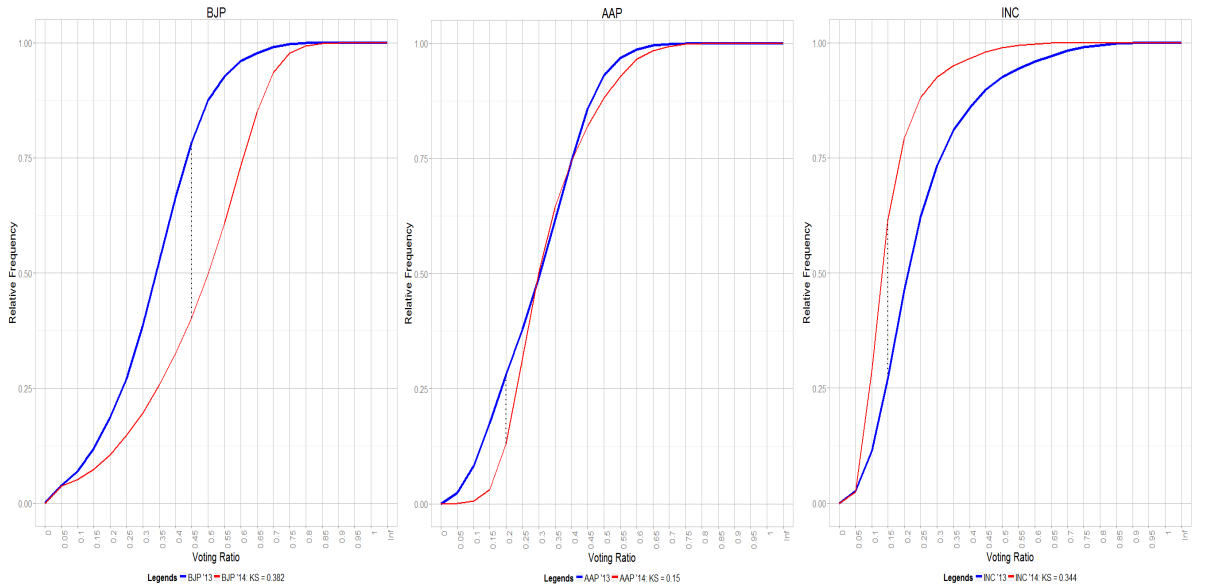

Fig. 3. KS test results for BJP, AAP and INC

5. CLUSTERING ANALYSIS

A possible way to understand voting transition is by first bucketing polling station into different categories and then analyzing the movement amongst these groups between the two elections. The standard way of clustering similar objects in statistics is known as K Mean Clustering technique. The idea behind the technique is to define a distance metric in the feature vector space of the objects and then clustering the objects which are close together. The most granular possible data is the polling booth level data and we have defined the feature vector based on the percentage votes achieved by the political parties. For example, suppose in a given polling station, BJP, AAP and INC secured v_1, v_2 and v_3 percentage votes. The feature vector for this polling station is denoted as (v_1, v_2, v_3). For all the polling stations in Delhi, similar vectors can be defined and plotted. We explain briefly *K-means clustering* algorithm employed to group similar behaving polling stations.

5.1 K-means algorithm

Given a set of observations $(x_1, x_2, , x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $S = S_1, S_2, , S_k$ so as to minimize the within-cluster sum of squares (WCSS) [7]. In other words, its objective is to find:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - u_i \|^2$$

The problem with K Means Clustering Algorithm is that it requires number of clusters as an input to the algorithm. In a two dimensional data, this can be found using visual inspection but for higher dimensions mathematical techniques should be used to define optimal number of clusters. In general, if there are n dominant parties in the state, there should be at least n clusters. Though there is a possibility of mixed clusters which may represent more than one party. In this work, an approach based on silhouette value has been used to identify the number of optimal clusters [6].

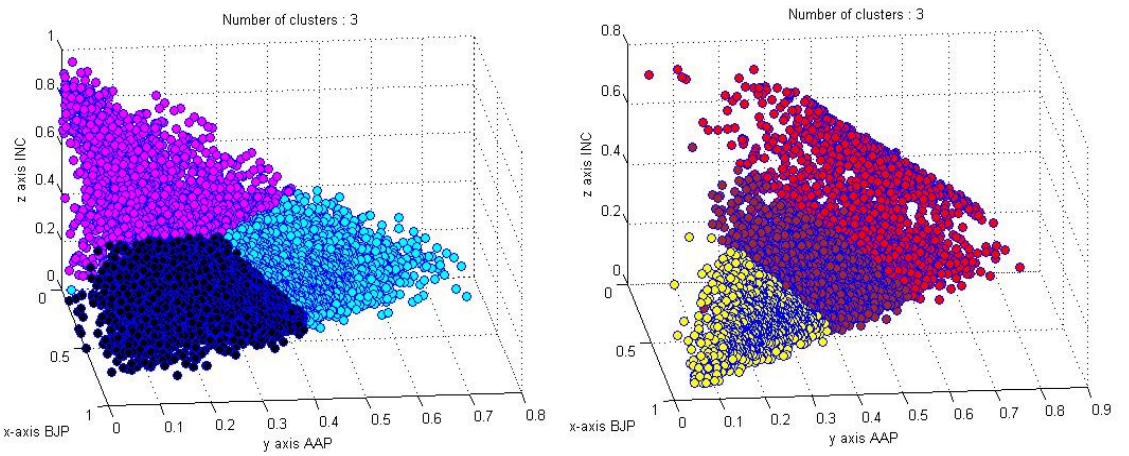

Fig. 4. Cluster plots in Legislative Election 2013 (L) and Parliamentary Elections 2014(R)

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the ith point, S(i), is defined as:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where a(i) be the average dissimilarity of i with all other data within the same cluster. a(i) is interpreted as how well i is assigned to its cluster (the smaller the value, the better the assignment). We then define the average dissimilarity of point i to a cluster c as the average of the distance from i to points in c.

Let b(i) be the lowest average dissimilarity of i to any other cluster which i is not a member. The cluster with this lowest average dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

For s(i) to be close to 1 we require a(i) $<<$ b(i). As a(i) is a measure of how dissimilar i is to its own cluster, a small value means it is well matched. Furthermore, a large b(i) implies that i is badly matched to its neighboring cluster. Thus an s(i) close to one means that the datum is appropriately clustered. If s(i) is close to negative one, then by the same logic we see that i would be more appropriate if it was clustered in its neighboring cluster.

The average s(i) over all data of a cluster is a measure of how tightly grouped all the data in the cluster are. Thus the average s(i) over all data of the entire dataset is a measure of how appropriately the data has been clustered. If there are too many or too few clusters, as may occur when a poor choice of k is used in the k-means algorithm, some of the clusters will typically display much narrower silhouettes than the rest. Thus silhouette plots and averages may be used to determine the natural number of clusters within a dataset.

5.2 Cluster Indentity

Using silhouette plots and averages, we find that no. of clusters in both Legislative Elections and Parliamentary Elections is three. On visualizing 3–D scatter plot (Fig. 4) of both the elections it is quite clear that composition of clusters

has changed in such a way so as to favor BJP.

Table V. No of polling stations won Assembly Elections 2013

Cluster 2013	AAP	AAP/INC	BJP	BJP/AAP	BJP/INC	INC	Cluster naming	% Polling Stations as per cluster naming
A	0	0	3905	0	3	19	**BJP**	99.4
B	5	0	8	0	1	1936	**INC**	99.28
C	3517	6	734	22	1	178	**AAP**	78.89

Suppose clusters formed in Legislative Elections are termed as A, B and C and clusters formed in Parliamentary Elections are termed as P, Q and R. We further associate each cluster with different parties using number of polling stations won by a particular party in that cluster. Table V and Table VI describes the identity of each cluster for both the elections. This analysis clearly shows how at poll booth level, BJP and AAP converted most of polls that belonged to Indian National Congress (INC) party.

Table VI. No of polling stations won in Parliamentary Elections 2014

Cluster 2014	AAP	AAP/INC	BJP	BJP/AAP	INC	Cluster naming	% Polling Stations as per cluster naming
P	1189	1	2374	16	57	AAP + **BJP**	32.69 & 65.27
Q	1610	1	0	0	254	**AAP**	86.32
R	0	0	4833	0	0	**BJP**	100

5.3 Movement of Polling Stations

Tagging of clusters with a party or party combination can be used to identify how polling stations have moved from one cluster to another. This analysis gives clear idea of how voting pattern changed over the course of few months (See Fig. 5).

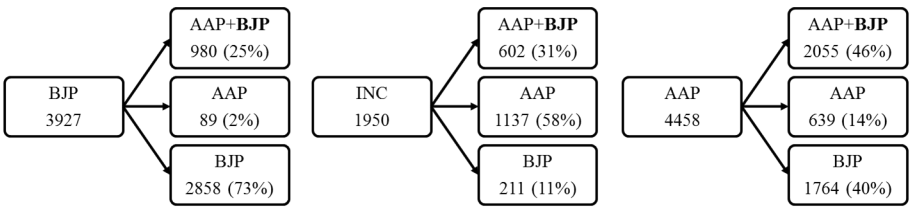

Fig. 5. Movement of polling stations from one cluster to another

6. REGRESSION MODEL FOR VOTE TRANSITION

The disadvantage of the clustering based approach is that it assumes a polling station to be a homogeneous entity. This assumption may be valid in a strong bipolar democracies like United States where the supporters of the two parties are clustered within an area and there is not much mobility between the two groups. In a multiparty democracy where more than one party represents the interests of a social group, the assumption of homogeneity within the same family may even be not true at times. Delhi election in 2013 and 2014 experienced a similar phenomenon and it is imperative to build a mathematical framework which does not have such a simplistic assumption.

The idea behind this approach is to analyze how voters in a polling station transitioned between political parties in the consecutive election. Consider a polling station which has n_1 registered electorate in the first election and n_2 registered electorate in the subsequent election. Even though these two numbers might not differ by much if the elections are held in a short span still for the sake of generality we have taken these two as different numbers. Suppose

in the first election, out of total n_1 voters, x_1 voted for BJP, x_2 voted for AAP, x_3 voted for INC, x_4 voted for other parties. Some of the registered voters might not have come to vote due to some reason and they can be denoted by x_5.

$$n_1 = x_1 + x_2 + x_3 + x_4 + x_5 \tag{1}$$

Similarly in the second election, out of total n_2 voters, y_1 voted for BJP, y_2 voted for AAP, y_3 voted for INC, y_4 voted for other parties. Some of the registered voters might not have come to vote due to some reason and they can be denoted by y_5.

$$n_2 = y_1 + y_2 + y_3 + y_4 + y_5 \tag{2}$$

The number of votes polled to BJP in the second election can be decomposed into various parts each of which coming from the voters of the different parties in the first election. For example some of the y_1 votes which BJP got in the second election were from the voters who had voted for BJP in the first election (x_1), some of them who had voted for AAP in the first election (x_2) and similarly from other groups. It should also be noted that some of the votes BJP got in the polling station might not be a registered voter in the first election, they can be donted as C_1.

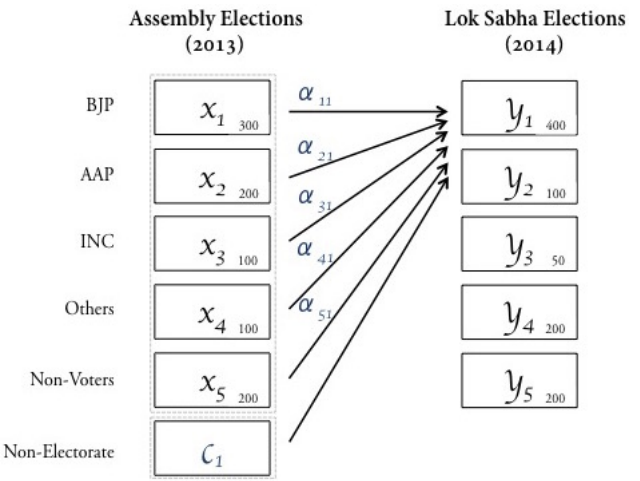

Fig. 6. Voter Transition Equation

We are interested in finding the percentage of voters from each group that voted for BJP in the second election. For example, α_{21} is proportion of voters that voted for AAP in the first election but transitioned to BJP in the second election. In general α_{ij}, is the proportion of voters of party i that switched to party j in the subsequent election. The model assumes that this transition coefficient would remain the same across all the booths. This assumption might be valid if the polling stations are with similar characteristics. Still with this assumption a general trend of voters in Delhi can be found which can be used to analyze the transition in voting pattern.

$$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \alpha_{13}x_3 + \alpha_{14}x_4 + \alpha_{15}x_5 + C_1 \tag{3}$$

Constrained ordinary least square regression is employed over the data-set to identify the vote transition. This quantifies the votes retained by each political party over the course of elections (Table II). As the transition coefficients can only be between 0 and 1, α_{ij} should be constrained as shown in the below equation.

$$0 \leq \alpha_{ij} \leq 1$$
$$\sum_{i=1}^{3} \alpha_{ij} \leq 1$$
$$\text{where } j = 1, 2 \textit{ and } 3$$

Expected number of votes for every poll booth can be estimated as per systems of equation mentioned below.

$$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \alpha_{13}x_3 + \alpha_{14}x_4 + \alpha_{15}x_5 + C_1$$
$$y_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \alpha_{23}x_3 + \alpha_{24}x_4 + \alpha_{25}x_5 + C_2$$
$$y_3 = \alpha_{31}x_1 + \alpha_{32}x_2 + \alpha_{33}x_3 + \alpha_{34}x_4 + \alpha_{35}x_5 + C_3$$

where y_1 represents vote count for BJP in Parliamentary elections, 2014 and similarly for AAP and INC. α_{ij} represents element of transition matrix. It also represents percentage of votes retained, for example, α_{11} represents percentage votes retained by BJP, α_{12},α_{13} and α_{14} represent percentage of votes acquired by BJP of AAP, INC and other political parties. α_{14} represents the contribution by those who did not vote in legislative assembly elections. C_1, C_2 and C_3 represents intercepts. These equations represent an optimization problem and can be solved using method *Generalized Reduction Gradient Algorithm* [4].

Table VII. Regression Results for all the polling stations

	2014 Results		**Initial Parameters**					
			BJP	**AAP**	**INC**	**Others**	**DNV**	**Intercept**
All 7 Parliamentary Constituencies	**BJP**	**Coeff**	1.00	0.43	0.00	0.12	0.00	9.82
	AAP	**Coeff**	0.00	0.42	0.46	0.44	0.08	9.06
	INC	**Coeff**	0.00	0.08	0.45	0.12	0.00	8.88
Chandani Chowk	**BJP**	**Coeff**	1.00	0.46	0.00	0.00	0.00	5.18
	AAP	**Coeff**	0.00	0.47	0.37	0.64	0.03	0.00
	INC	**Coeff**	0.00	0.03	0.54	0.25	0.00	8.12
North East Delhi	**BJP**	**Coeff**	1.00	0.51	0.00	0.08	0.03	0.00
	AAP	**Coeff**	0.00	0.49	0.54	0.42	0.05	0.00
	INC	**Coeff**	0.00	0.00	0.44	0.27	0.00	4.60
East Delhi	**BJP**	**Coeff**	1.00	0.58	0.00	0.00	0.00	18.06
	AAP	**Coeff**	0.00	0.42	0.44	0.63	0.04	20.83
	INC	**Coeff**	0.00	0.00	0.54	0.20	0.04	2.22
New Delhi	**BJP**	**Coeff**	0.94	0.49	0.00	0.00	0.00	3.28
	AAP	**Coeff**	0.06	0.36	0.47	0.94	0.02	0.00
	INC	**Coeff**	0.00	0.14	0.45	0.06	0.03	14.82
North West Delhi	**BJP**	**Coeff**	0.99	0.39	0.00	0.19	0.00	0.00
	AAP	**Coeff**	0.00	0.50	0.59	0.49	0.08	0.00
	INC	**Coeff**	0.00	0.12	0.31	0.06	0.01	0.00
West Delhi	**BJP**	**Coeff**	0.96	0.36	0.00	0.37	0.00	7.36
	AAP	**Coeff**	0.04	0.43	0.35	0.14	0.13	0.00
	INC	**Coeff**	0.00	0.12	0.38	0.08	0.05	0.00
South Delhi	**BJP**	**Coeff**	0.96	0.11	0.00	0.17	0.16	4.55
	AAP	**Coeff**	0.00	0.65	0.51	0.51	0.00	0.00
	INC	**Coeff**	0.00	0.15	0.27	0.05	0.00	8.62

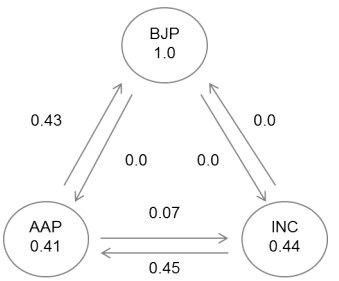

Fig. 7. Voting transitions in all Parliamentary Constituencies combined

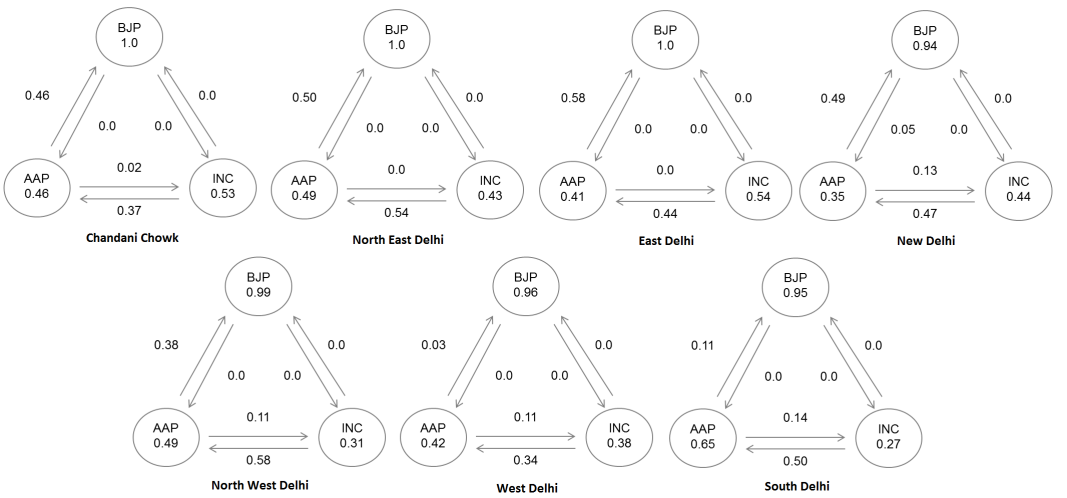

Fig. 8. Voting transitions in all Parliamentary Constituencies combined

It can be seen from the results that while BJP retained its voters, it gained from many of the AAP voters of 2013 elections (See Fig. 7 and 8). Even though the primary analysis shows that AAP retained its vote share, regression based model validates the popular perception that AAP lost its major voter base in the second election. There was no net reduction in AAP votes, as it encroached the voter base of INC, This can also be substantiated from the fact that AAP pursued left of center policies aimed at the poorer sections which have been the traditional vote bank of INC.

7. CONCLUSION

This paper develops a statistical framework to measure the transition in voting pattern in consecutive elections. The work shows that traditional mean median kind of analysis is insufficient to identify the movement of votes. It recommends to use most granular data possible in the analysis which for India consists of polling station level data. It builds a K mean clustering algorithm to bucket different polling stations and to use them in identifying the movement in voting pattern. This methodology may work for cases where there is homogeneity within a polling station. In a multi party democracy there are various parties competing for the same social group, the homogeneity assumption may not always work. As a results a constrained linear regression based approach can be used which can provide very meaningful results.

The results obtained from the constrained linear regression approach can be utilized to understand the effect of political strategy taken by the party. For example, the results clearly show that the strategy of AAP to make inroads in INC vote bank miserably back fired as most of its support went to BJP. Even though the loss of voters to BJP were compensated by the voters from INC, ot strengthened the prime challenger of AAP. In a first past the post system, this resulted in a complete domination of BJP. Similar analysis can be done for other political scenarios so that such strategies are not repeated. Such political intervention based on statistical analysis can be extremely useful in the political scene. Some of the techniques employed in this analysis like the clustering technique, can help the pollsters to construct better survey design. This can be very helpful in designing a representative survey which would be able to correctly predict elections using optimal number of survey respondents.

REFERENCES

D. Avello, P. Metaxas, and E. Mustafaraj. Limits of electoral predictions using twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

A. Bermingham and A. Smeaton. On using twitter to monitor political sentiment and predict election results. *Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011*, pages 2–10, 2011.

A. Bonica. Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386, 2014.

R.H Byrd, J.C Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89(1):149–185, 2000.

ECI. Election commission of india, delhi office. http://ceodelhi.gov.in/Content/pastelection.aspx, 2014.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. ISBN: 978-0-471-73578-6. Wiley Series in Probability and Statistics, 1990.

S. Llyod. Least squares quantization in pcm. *Journal of Statistical Software*, 28(18):129–137, 1982.

G. Marsaglia, W. Tsang, and J. Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18), 2003.

F.J Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

L.H Miller. Table of percentage points of kolmogorov statistics. *Journal of the American Statistical Association*, 51(273):111–121, 1956.

Brid-Aine Parnell. Scientists warn about bias in the facebook and twitter data used in millions of studies. http://www.forbes.com/sites/bridaineparnell/2014/11/27/scientists-warn-about-bias-in-the-facebook-and-twitter-data-used-in-millions-of-studies/, 2014.

T. Rusch, I. Lee, K. Hornik, W. Jank, and A. Zeileis. Influencing elections with statistics: Targeting voters with logistic regression trees. *The Annals of Applied Statistics*, 7(3):1612–1639, 2013.

Karthik Shashidhar. Why it's difficult to have a nate silver in india. http://www.livemint.com/Politics/Rzc3n0qfBHe35G00bbRSpM/Why-its-difficult-to-have-a-Nate-Silver-in-India.html, 2013.

Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, and Jacob Spoelstra. Predicting us primary elections with twitter, 2012.

Jon Wakefield. Ecological inference for 2x2 tables. *Royal Statistical Society*, 167(3):385–445, 2004.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and us ers from user votes. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.