computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
On-line ISSN 2312-5381
International Journal of Computing

IMPROVING SUPPORT ESTIMATES BY FUSION OF PRE-ELECTION DATA

Miki Sirola [1)], Jaakko Talonen [1)], Mika Sulkava [2)]

[1)] Aalto University, P.O. Box 15400, 00076 Aalto; Finland,
Miki.Sirola@aalto.fi, https://users.ics.aalto.fi/miki/, Jaakko.Talonen@gmail.com, http://jaakko.me
[2)] Natural Resources Institute Finland (Luke), Statistical services, P.O. Box 1003, FI-00581 Helsinki, Finland,
mika.sulkava@luke.fi, www.luke.fi/en/personnel/mika-sulkava

Abstract: In this paper, Gallup results and a questionnaire in the context of a voting advice application related to the Finnish presidential election are combined. The main emphasis is on preprocessing phases where raw data is reformed to temporal data sets. We also pay attention to find optimized parameters for a merged recursive model. Aggregated data from a questionnaire was stored frequently and modified by a differential equation. The method presented in this paper allows us to visualize more accurately the daily support of each candidate before the election. The results can be used for further research such as forecasting the results and the success of presidential campaigns. *Copyright © Research Institute for Intelligent Computer Systems, 2015. All rights reserved.*

Keywords: election; time series; data fusion.

1. INTRODUCTION

In this paper, we propose methods for combining two different data sources to get a more accurate estimate of election candidates' support in time. Repeated surveys are drawn from a population at irregular time intervals during the campaign. The surveys are reformed into time series with regular time intervals. The preprocessed data are combined with a second data source: questionnaire answers in the context of a voting advice application (VAA). VAAs are increasingly popular in democracies worldwide, especially among a group that is often considered 'apathetic' about electoral politics: youth [1]. Wrong Gallup designs, as in webpage VAAs, are producing sample proportions that differ systematically from the population [2]. Although, with preprocessing and parameter optimization VAA data can be used to make the support estimates more accurate.

Sampling over time by repeated polls enables analysis of process through estimation [3-5]. In addition to the usual design issues we need to consider the frequency of sampling, the spread and pattern of inclusion of units over time, the use of overlapping or non-overlapping samples over time, and the precise pattern of overlap. Time series is produced from repeated surveys by estimating the number of interviewed people per day. The analysis of these time series may involve seasonal adjustment and trend estimation [6]. In this paper, we use well known simple time series modeling techniques, but on the other hand, we introduce a new empirical method to combine two different types of repeated surveys series together. This article is an extended version of the original paper [7].

2. RELATED WORK

Political decision making is studied quite a lot in literature, e.g. in [8] and [9]. The articles discuss the topic in many countries around the Europe and the whole world. Prediction of election results is one quite popular issue.

Web-based theoretical studies and survey-based studies are discussed in [10]. Reflections on the discrepancies between mobilization and normalization of political participation are offered.

In [11] take-the-best heuristic is used to develop a model to forecast the popular two-party vote shares in U.S. presidential elections. The model draws upon information about how voters expect the candidates to deal with the most important issue facing the country. The model forecasts were competitive with the forecasts from methods that incorporate much more information.

Voting advice applications are studied also in [12], where Belgian voters are analyzed. It was noticed that different combinations of statements produced diverging information for the participants.

Sometimes the statements have remarkable impact on the 'voting advice' produced.

An election campaign in Belgium is discussed also in [13], where the electoral impact of a popular VAA and TV show are empirically assessed. The study systematically compares the differences in voting behavior in a large panel of Internet users. According to the test, the VAA affected to the Belgian voters final decision.

The distortion of Gallup polls in various countries is discussed in [14]. The polls have a tendency to exaggerate the changes during the campaign compared with the Election Day results and the situation in the beginning of the campaign. One explanation is claimed to be people getting more responsible for their views when stepping in the voting booth.

The behavioral sciences try to re-evaluate the current practice of relying on single sample null hypothesis tests [15]. U.S. Presidential election is studied as an example. More accurate prediction was obtained with single samples compared to polling percentages. Changes in the statistical approaches in psychology have been proposed after improved predictions.

A strategy to properly analyze old public opinion data is composed in [16]. A large set of eighty-year-old American public opinion data is studied. Quota-control methods are used to solve the problems associated with this kind of data. The aim is to improve the data-analysis both in aggregate and individual level. The methods of analysis laid out in the article enable to utilize historical public opinion data.

Relationship among three contemporary concepts conflated in the literature: anti-establishment politics, political outsiders and populism are clarified in [17]. Empirical studies on representation are mainly based on descriptive levels of political and ideological congruence between parliament members and voters [18]. Very few studies focus on explaining congruence, or if they do the explanatory dimensions are too spare.

The creation of terrorist organizations by political parties is discussed in [19]. Institutional structural constraints are noticed to increase such a risk according to literature. The party ideology has also been claimed to be an important factor to correlate with terror tendency. A supervised machine-learning approach and data from mono-thematic Twitter accounts are used in figuring out political activism, especially considering separation of political activists and general public in social media conversation [20].

The importance of party type is tested in [21] for an explanation of levels of intra-party congruence. The test is controlled by using the main congruence-assessment methods. The focus here is in Portuguese party system.

Voting advice applications data is often used to test many empirical questions regarding voting behavior and political participation, but fewer approaches exist to use VAAs to estimate the positions of political parties. In [22] methodological issues regarding the phasing of statements, the format of response scales, the reliability of coding statements into response scales and the reliability and validity of scaling items into dimensions are examined.

In [23] the main dimensions of competition between parties in Romania are analyzed by surveying patterns of the party in the electorate. The reliability of using VAA data for Romanian party mappings is examined and the results are compared with other party mappings. The determinants of positional incongruence between pre-election statements and post-election behavior in the Swiss parliament are examined in [24].

This literature review views the studies of political decision making in general and methodological level. Deeper methodological analysis is left for future work. The survey goes through the related work to help to better understand the location of this work in the larger entity in the field of political science.

3. DATA

The president of Finland is elected by a direct election for a term of six years according to the constitution. The most recent Finnish presidential election was held in January-February 2012. The president is elected in two rounds if necessary. This paper focuses on the time before the first election round, which was held on Sunday 22 January 2012. The former president, Tarja Halonen, was no longer able to stand for a third consecutive term. In presidential elections all Finnish citizens who meet the 18 years before election day can vote. About 4.4 million people in Finland have a possibility to vote [25].The candidates and their accession numbers, parties, and abbreviations are shown in Table 1.

Table 1. The candidates, their accession numbers, and parties.

№	Candidate	Party
9	Paavo Arhinmäki (PA)	Left-Wing Alliance (VAS)
8	Eva Biaudet (EB)	Swedish People's Party in Finland (PKP)
7	Sari Essayah (SE)	Christian Democrats in Finland (KD)
2	Pekka Haavisto (PH)	Green League (VIHR)
5	Paavo Lipponen (PL)	The Finnish Social Democratic Party (SDP)
6	Sauli Niinistö (SN)	National Coalition Party (KOK)
3	Tino Soini (TS)	The Finns (PS)
4	Paavo Väyrynen (PV)	Centre Party of Finland (KESK)

In our research two main data resources were used. First data set was captured from Helsingin

Sanomat (HS) VAA [26]. HS VAA provides a channel for candidates to tell their opinions to various topical issues. Questions from different topics were available for the candidates to answer in advance. The voters express their views on these issues. The output of the VAA is a ranked list of candidates.

In the beginning of December 2011 HS released the data of the candidates' and users' answers [27]. HS VAA had also a question, where it was asked which candidate the user will vote. This information was not included in the published data, neither was the best-matching (BM) candidate algorithm open. However, total statistics of each candidate's support and sample size was shown at [26]. We collected the candidate support values every hour, but the sample sizes were manually collected with varying frequency; see the left side of Fig. 1. Data collection was started about two months before the elections.

Second data resource was Gallup results provided by different research institutes, see the right side of Fig. 1. All Gallup results were collected and downloaded from Wikipedia [28]. Other data sources are the results of different polls. These results are planned to be used in further research such as election result forecasting.

3.1. GALLUPS

The most reliable data used in this paper are 17 Gallup studies. The studies are mostly computer-assisted telephone interviews. The interviews had a target group from 18 to 79-year-old population, the Åland Islands excluded. The sample was formed randomly from the population information system. The sample is weighted to the population by age, sex and place of residence. It is expected that the interviews are performed same way in other Gallups (TNS-Gallup, Research Insight Finland, MC-Info) too [29].

The margin of error consequent of selected sample is well defined. The results in the second round indicate that sampling was performed carefully. Five Gallups were done during two weeks between the first and second election round. Sauli Niinistö got 65, 64, 64, 63 and 62% support in these polls. In the election 1802400 voted for him and he got 62.60% of votes [28].

In parliamentary elections and in the previous presidential elections, the support percentages were calculated by using only those voters who knew their candidate. A new way to publish Gallup results includes "does not know" and "did not respond" (in this paper "non") groups.

It involves that the sum of support of all candidates is less than 100%.

Interview periods and sample sizes varied between the Gallups, see Table 2. The reason for this is naturally different poll sources. Typically these are not taken into account in analyses. E.g., in [30] each poll has a constant weight. In [31] the interval between polls is misleadingly visualized as constant. Of course, for the first analysis the accuracy of this visualization is sufficient. Better results are achieved if the results are weighted by the sample sizes.

Table 2. Gallup date, source, sample size, and support numbers. Source abbreviations: Taloustutkimus / YLE (TT), TNS-Gallup / HelsinginSanomat (TNS), MC-Info / Ilta-Sanomat (MC) and Research Insight Finland / Iltalehti (RIF). A non-value is shown in parentheses indicates that in these Gallups support percentages were calculated only from those who gave answers to the poll [28].

Date	Source	Sample	PA	EB	SE	PH	PL	SN	TS	PV	non
10.-18.1.2012	TT	1457	4	2	2	12	5	29	6	10	30
11.-15.1.2012	TNS	1408	6	2	3	17	6	39	9	17	(31)
9.1-13.1.2012	RIF	1014	3	2	2	11	3	37	7	12	22
9.-11.1.2012	MC	1000	7	2	3	12	5	49	9	13	(18.5)
19.12.2011-3.1.2012	TT	1464	4	2	1	8	4	37	7	8	29
21.-29.12.2011	RIF	1011	3	1	2	8	4	41	9	11	19
13.-30.12.2011	TNS	1473	4	2	2	7	7	38	9	9	21
15.-20.12.2011	MC	1000	4	3	3	9	9	51	11	11	(18)
30.11.-14.12.2011	TT	1685	3	3	2	6	5	40	7	9	25
29.11.-12.12.2011	TNS	1979	5	4	2	6	6	43	11	9	14
15.-26.11.2011	TNS	978	4	3	2	5	7	41	9	8	21
14.-25.11.2011	RIF	1000	3	3	1	6	6	43	6	7	26
9.-17.11.2011	TT	1452	3	3	1	6	7	49	8	8	15
31.10.-13.11.2011	TNS	<1000	3	3	1	6	7	44	11	10	14
24.10.-4.11.2011	RIF	>1000	1	2	1	6	5	47	9	6	23
4.-15.10.2011	TNS	980	2	3	2	6	7	50	8	6	16
26.-29.9.2011	RIF	1010	2	2	1	6	11	49	11	6	15

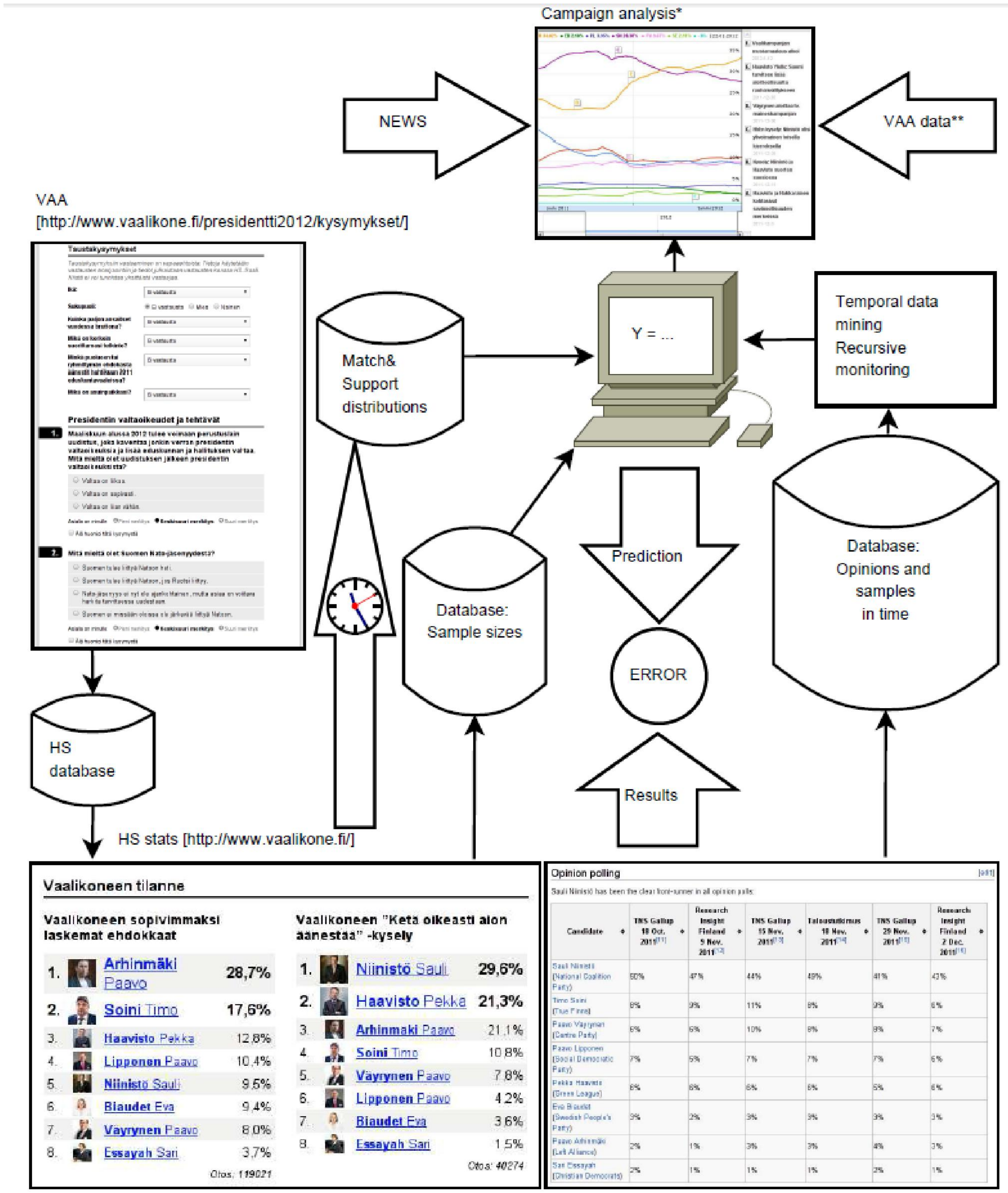
In our analysis it is expected that the sampling frequency in each Gallup is constant between the start and end dates. In addition it is possible to weight surveys depending on the poll source. In this paper, it was expected that each source i is as reliable as the others (w_i=1, $\forall i$). The survey results were combined using a simple recursive algorithm.

Recent Gallup results were weighted more and, e.g., polls (part of the Gallup) one week earlier were weighted with λ^7.

3.2. WEBSITE POLL

In addition to the VAA, HS had a poll "who you will vote?" in their webpage.More than 400000 people answered it and about 25% gave also background information [26]. Since the end of November until the election day of the first round, the proportions of the candidates suggested by HS VAA as best matches and cumulative support data were stored every hour. Cumulative "who you will vote?" statistics are shown in Fig. 2.

Useful data is not always available for a researcher, but current sample sizes and percentages were shown in the webpage during the presidential campaign. Therefore, data was stored automatically. It was possible to store the total support of each candidate every hour, and sample sizes were stored manually a few times.



Taustakysymykset

Taustakysymyksiin vastaaminen on vapaaehtoista. Tietoja käytetään vastausten analysointiin ja tiedot julkaistaan vastausten kanssa vain ... Niistä ei voi tunnistaa yksittäisiä vastaajia.

Ikä:	Ei vastausta
Sukupuoli:	● Ei vastausta ○ Mies ○ Nainen
Kuinka paljon ansaitset vuodessa brutona?	Ei vastausta
Mikä on korkein suorittamasi tutkinto?	Ei vastausta
Minkä puolueen tai ryhmittymän ehdokasta äänestit huhtikuun 2011 eduskuntavaaleissa?	Ei vastausta
Mikä on asuinpaikkasi?	Ei vastausta

Presidentin valtaoikeudet ja tehtävät

1. Maaliskuun alussa 2012 tulee voimaan perustuslain uudistus, joka kaventaa jonkin verran presidentin valtaoikeuksia ja lisää eduskunnan ja hallituksen valtaa. Mitä mieltä olet uudistuksen jälkeen presidentin valtaoikeuksista?

○ Valtaa on liikaa.
○ Valtaa on sopivasti.
○ Valtaa on liian vähän.

Asialla on minulle ○ Pieni merkitys ● Keskinkertainen merkitys ○ Suuri merkitys

□ Älä huomioi tätä kysymystä

2. Mitä mieltä olet Suomen Nato-jäsenyydestä?

○ Suomen tulee liittyä Natoon heti.
○ Suomen tulee liittyä Natoon, jos Ruotsi liittyy.
○ Nato-jäsenyys ei nyt ole ajankohtainen, mutta asiaa on voitava harkita tarvittaessa uudestaan.
○ Suomen ei missään oloissa ole järkevää liittyä Natoon.

Asialla on minulle ○ Pieni merkitys ● Keskinkertainen merkitys ○ Suuri merkitys

□ Älä huomioi tätä kysymystä

Vaalikoneen tilanne

Vaalikoneen sopivimmaksi laskemat ehdokkaat

1.	Arhinmäki Paavo	28,7%
2.	Soini Timo	17,6%
3.	Haavisto Pekka	12,8%
4.	Lipponen Paavo	10,4%
5.	Niinistö Sauli	9,5%
6.	Biaudet Eva	9,4%
7.	Väyrynen Paavo	8,0%
8.	Essayah Sari	3,7%

Otos: 119021

Vaalikoneen "Ketä oikeasti aion äänestää" -kysely

1.	Niinistö Sauli	29,6%
2.	Haavisto Pekka	21,3%
3.	Arhinmäki Paavo	21,1%
4.	Soini Timo	10,8%
5.	Väyrynen Paavo	7,8%
6.	Lipponen Paavo	4,2%
7.	Biaudet Eva	3,6%
8.	Essayah Sari	1,5%

Otos: 40274

Opinion polling [edit]

Sauli Niinistö has been the clear front-runner in all opinion polls:

Candidate	TNS Gallup 18 Oct. 2011[11]	Research Insight Finland 5 Nov. 2011[12]	TNS Gallup 15 Nov. 2011[13]	Taloustutkimus 18 Nov. 2011[14]	TNS Gallup 29 Nov. 2011[15]	Research Insight Finland 2 Dec. 2011[16]
Sauli Niinistö (National Coalition Party)	50%	47%	44%	49%	41%	43%
Timo Soini (True Finns)	8%	9%	11%	8%	9%	6%
Paavo Väyrynen (Centre Party)	6%	6%	10%	8%	8%	7%
Paavo Lipponen (Social Democratic Party)	7%	5%	7%	7%	7%	6%
Pekka Haavisto (Green League)	6%	6%	6%	6%	5%	6%
Eva Biaudet (Swedish People's Party)	3%	2%	3%	3%	3%	3%
Paavo Arhinmäki (Left Alliance)	2%	1%	3%	3%	4%	3%
Sari Essayah (Christian Democrats)	2%	1%	1%	1%	2%	1%

Fig. 1 – Data mining phases. Left) Citizens fill VAA and answers are stored to HS database. Aggregated data is captured from the website and stored in our database every hour. Right) A collection of published Gallup data sets are stored in our database. Computer) The data sets are combined and analyzed. The results are used for election results prediction. *) Merged data can be used for analyzing the campaign. Hindsighting/media monitoring was shortly done in this paper without any political expert help

Sample sizes were stored manually 32 times with random frequency during this time. Missing sample size values were estimated by combination of linear and moving average (MA) methods. The number of missing sample values is large, because the sample sizes were stored manually. The difference of sample size is always positive and approximately constant. Therefore a simple combination of a linear algorithm and MA was performed. The output of this method is hourly approximation of cumulative sample size of VAA users. In the experiments, these cumulative values are reformed into averaged support values. Solutions for several problems such as modifying the sum of support percentages to 100%, combining sample sizes and rounded percentages, etc. are explained in the experiments.

4. METHODS FOR GALLUP PROCESSING

4.1. GALLUP PROCESSING

Published Gallup support values are rounded to the nearest integer as can be seen from Table 2.

Candidates with low support have larger proportional error. E.g., candidate SE has 5-14 supporters in the RIF poll organized on 14th-25th of November (the sample size was 1000).

There exist two different types of polls. In the first type, the sum of candidates' support (including or excluding non) is about 100%. In the second type in three Gallups, the sum of candidates' support is 100%, so those numbers are calculated only from poll respondents who had chosen a candidate or told it.

Accurate poll data was not available, but with a few preprocessing phases accuracy of integer results can be improved. Preprocessing phases for Gallup data are shown in Fig. 3. Phase functions *Fi* (processes) and questions *Qi* (decisions) are:

F1 Gallup type change: each support percentage X_i is multiplied by a coefficient $X_i = X_i \cdot (100-X_{non})/100$.

F2 Scaling: $X_i=X_i-(\sum_i^N X_i+X_{non}-100)/(N+1)$, where N is the number of candidates.

F3 If a Gallup does not have an exact sample size, new sample size is calculated as $S^*=\sum X_{2,i} \cdot S$.

F4 If S^* is larger than the Gallup sample size, the number of candidate supporters S_u^* in the Gallup is reduced by one, where u is $\min(S_i^*-\lfloor S_i^*+0.5\rfloor)$.

F5 If S^* is less than the Gallup sample size, the number of candidate supporters Su^* in the Gallup is increased by one, where u is $\max(S_i^*-\lfloor S_i^*+0.5\rfloor)$.

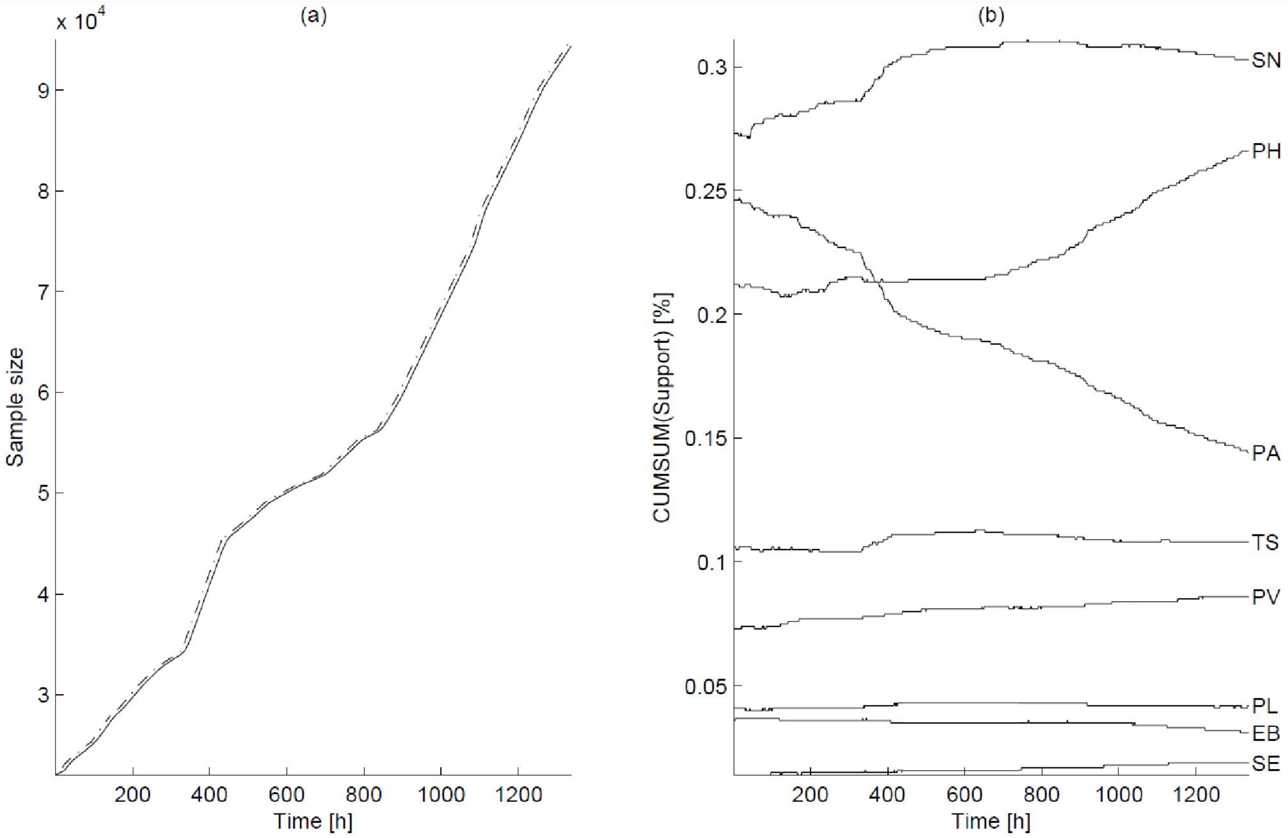


Fig. 2 – Stored "who you will vote" -data: (a) cumulative sample size (dotted) and MA (solid line). (b) the candidates' cumulative support in time

exit The loop is forced to exit, e.g., in a situation where S=S^*+1 or S=S^*-1 and $|u|$=2 in phases F4 and F5. In this case Gallup improvement is failed and decimal fractions for S_i^* are used to ensure same S^* as in the original data.

Q1 Traditional Gallup or the version where $\sum S+non\approx 100$?

Q2 Is accurate Gallup sample size S_i available? See sample sizes of each Gallup in Table 2.

Q3 Is the total number of supporters the same as original sample size $S^*=\sum_i^N\lfloor X\cdot S+0.5\rfloor=S$?

Q4 Is the sum of scaled and rounded sample sizes larger than the original sample size, $\sum_i^N S_i^{\,*}+S^*_{non}>S$?

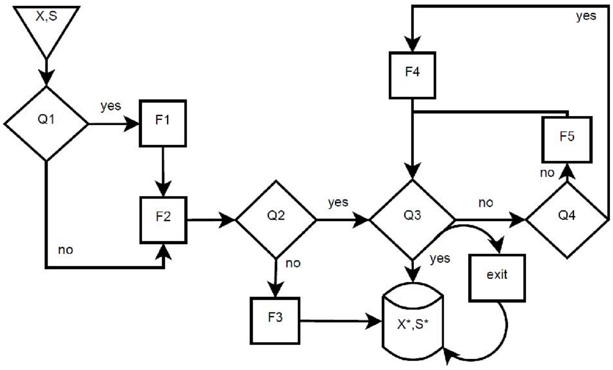


Fig. 3 – Gallup preprocessing flow chart. Support value matrix *X* and sample size *S* are inputs. Outputs are modified matrices including information of all candidate support values and sample sizes, see Table 3

4.2. ADDING SAMPLE SIZE TO DATE INFORMATION

Each Gallup has a start and an end date. As mentioned before, these studies are mostly computer-assisted telephone interviews, so it can be assumed that the sample size is uniformly distributed between these dates. Active Gallups matrix *T* is defined as

$$\mathbf{T}=\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}, \tag{1}$$

where $x_{t,j}$ is zero or one and n is the length of the whole data set (from the start date of the first Gallup until the elections) and k is the number of all surveys in this period.

The elapsed time of each poll is defined as

$$L_j=\sum_{t=1}^{n}\mathbf{T_{t,j}}, \tag{2}$$

The number of interviews for each Gallup *j* per each day *t* is estimated as

$$\mathbf{N_{t,j}}=\frac{\mathbf{T_{t,j}}\cdot S_j}{L_j}. \tag{3}$$

where S_j is the total sample size of Gallup *j*.

4.3. CANDIDATE SUPPORT IN TIME

Let y_t represent the standard sample survey estimate of θ_t, based on the sample at time *t*. We may define,

$$y_t=\theta_t+e_t, \tag{4}$$

where error term e_t is zero mean and $var(e_t)=s_t^2$ [3].

Observations are not made at equally spaced time intervals. Therefore matrix *N* is defined, where observations, see (4) are estimated/generated between the samples (Gallups) by a simple first-order autoregressive model. By a recursive algorithm [32] the weighted matrix N^* for supporters of candidate *i* is defined as

$$N^*_{t,j}=\lambda\cdot N^*_{t-1,j}+N_{t,j}+e_{t,j}, \tag{5}$$

where the past numbers of supporters are weighted with a forgetting factor $0<\lambda\leq 1$. We are using this simple first-order autoregressive model for estimating the Gallup support values. The main emphasis will be on combining this model to VAA estimates.

The support of candidate *i* (including $\mathrm{Y}_{\mathrm{non}}$) matrix *Y* is defined as

$$\mathbf{Y_{t,i}}=\frac{\sum_{j=1}^{k}\left(N^*_{t,j}\cdot X_{j,i}\right)}{\sum_{j=1}^{k}N^*_{t,j}}, \tag{6}$$

where *Y* is $(n\times(N+1))$, *N* is the number of candidates, and $X(j,i)$ is the support of candidate *i* in Gallup *j*.

5. METHODS FOR VAA PROCESSING

Stored raw data from the website poll is reformed before combining it with Gallup data. Candidate data in time is estimated by cumulative percentage data $C(t)$, cumulative sample sizes $s(i)$ and recording time $t(i)$, $i\in[1,\ldots,k]$, where *k* is the number of data collected manually.

As in published Gallup values, VAA values are rounded, but in this case to the nearest first decimal. Candidates with low support have larger proportional error. Because of this inaccuracy $C(t)$ rescaled so that $\sum_i^N C_{i,t}=100.0, \forall t$.

Hourly cumulative sample size $S(t)$ is estimated by linear interpolation between two known points. I.e., we find approximate values of a function $S(t)$[p.936-957] [33]. This is a special case of polynomial interpolation with $n=1$. Obviously, there are other kinds of approximations, but in our research we found it practical to avoid "sharp angles" using a MA function to $S(t)$.

The cumulative number of supporters for candidate *i* is defined as $X_i(t)=C_i(t)\ S_i(t)$. Inaccuracy in $C(t)$ can cause reduction in function $X_i(t)$. It is corrected so that $X_i(t\text{-}1)\leq X_i(t), \forall t$. The estimated amount of supporters of candidate i in current time is defined as

$$\mathbf{V_{t,i}} = \frac{X_i(t) - X_i(t-a)}{S_i(t) - S_i(t-a)}, \qquad (7)$$

where a is the size of the sliding window. In this paper, preprocessed V values are used in further analysis.

5.1. COMBINING TWO DATA SOURCES

The main goal in this paper is to introduce a strategy to combine matrices Y and V, see (6) and (7). The support percentages based on VAA (matrix V) are biased, so the first difference of it is used. Also the frequency of V and Y can be different (hours and days).

The sample sizes of VAA and combined Gallup information varies in time, so a new time depended coefficient $\gamma(t)$ is defined as

$$\gamma(t) = \alpha \cdot \frac{D(S(t))}{\sum_i N^*(t,i)}, \qquad (8)$$

where $D(S(t))$ is the sample size of VAA and N^* is the number of supporters in current time t, see (5).The coefficient varies on every time step (day) depending on the proportion of data source sample sizes.

Combined matrix Z is defined as

$$\mathbf{Z(t,i)} = \mathbf{Y(t,i)} + \gamma(t) \cdot D(\mathbf{V(t,i)}) \qquad (9)$$

and

$$\mathbf{Z(t,N+1)} = \mathbf{Y(t,N+1)}\,. \qquad (10)$$

After combining matrices, vectors $Z(t,:)$ are rescaled so that $\sum_i^{N+1} Z_t, i$=100.0, $\forall t$.

5.2. PARAMETER OPTIMIZATION

The Mean Square Error (MSE) is an estimator of the overall deviations between modeled X_i and measured values divided by sample size of the real measured signal outputs T. In this paper, parameter α in the time dependent coefficient $\gamma(t)$ for VAA and forgetting factor λ for merged Gallup result were optimized. The MSE of each Gallup result one day earlier was minimized.

The error matrix ε is defined as

$$\varepsilon_{MSE(\alpha,\lambda)} = \frac{\sum_{i=1}^{N-1} (X_i(t_{start}-1,:) - \hat{X}_i(t_{start}-1,:))^2}{(N-1)-1}, \qquad (11)$$

where N is the number of Gallups. There is no estimation before the first Gallup results and, therefore, N-1 is used.

ε, X and $\hat{H}_i$ are N-1 ×n_C + 1 matrices. The error for each Gallup is defined by

$$\varepsilon_G = \frac{\sum_{i=1}^{n_C+1} \varepsilon_{MSE(\alpha,\lambda)}}{n_C + 1}, \qquad (12)$$

where n_C is the number of candidates and error for each candidate and C_{non} is defined by

$$\varepsilon_C = \frac{\sum_{i=1}^{n_G} \varepsilon_{MSE(\alpha,\lambda)}}{n_G}. \qquad (13)$$

6. EXPERIMENTS

The VAA data was stored every hour, but the Gallup result interval is 24 hours. Matrix V (7) was calculated and the results are shown in Fig. 4. The daily VAA support was simply calculated by averaging $V_{t_{day}} = \left(\sum V_{t_h \cdots t_{h+23}}\right) / 24$. Preprocessing for each Gallup was performed as shown in Fig. 3 resulting in new candidate support estimates, see Table 3.

Table 3. Preprocessed Gallup results. Start and end dates t, sample sizes S with and without non, and corrected support numbers. *) approximated.

$t_{i,start}$	$t_{i,end}$	source	$S_i \bigcap non_i$	$S_i \setminus non_i$	PA	EB	SE	PH	PL	SN	TS	PV	non
1	4	RIF	1010	862	1.68	1.68	0.69	5.64	10.69	48.61	10.69	5.64	14.65
9	20	TNS	980	823	2.04	2.96	2.04	6.02	6.94	50	7.96	6.02	16.02
29	40	RIF	*1053	811	1.04	1.99	1.04	5.98	5.03	46.91	9.02	5.98	22.98
36	49	TNS	*952	818	3.15	3.15	1.16	6.09	7.14	44.01	11.13	10.08	14.08
45	53	TT	1452	1234	3.03	3.03	0.96	5.99	7.02	48.97	7.99	7.99	15.01
50	61	RIF	1000	741	2.9	2.9	0.9	5.9	5.9	42.8	5.9	6.9	25.9
51	62	TNS	978	773	3.99	2.97	2.04	5.01	7.06	41	9	7.98	20.96
65	78	TNS	1979	1702	5	3.99	1.97	6.01	6.01	43	11.02	8.99	14
66	80	TT	1685	1264	3.02	3.02	2.02	5.99	4.98	39.98	7	9.02	24.97
81	86	MC	1000	821	3.2	2.4	2.4	7.3	7.3	41.7	8.9	8.9	17.9
87	95	TNS	1011	819	2.97	0.99	1.98	8.01	6.03	41.05	9	10.98	18.99
79	96	RIF	1473	1162	4.14	2.1	2.1	7.13	7.13	39.08	9.1	9.1	21.11
85	100	TT	1464	1039	4.03	1.98	1.02	7.99	4.03	37	6.96	7.99	29.01
106	108	TNS	1000	815	5.7	1.6	2.5	9.8	4.01	39.9	7.3	10.6	18.5
106	110	RIF	1014	790	3.16	2.07	2.07	11.14	3.16	37.08	7.1	12.13	22.09
108	112	TNS	1408	970	4.19	1.49	2.13	11.79	4.19	26.99	6.32	11.79	31.11
107	115	RIF	1457	1020	3.98	1.99	1.99	12.01	5.01	29.03	5.97	10.02	29.99

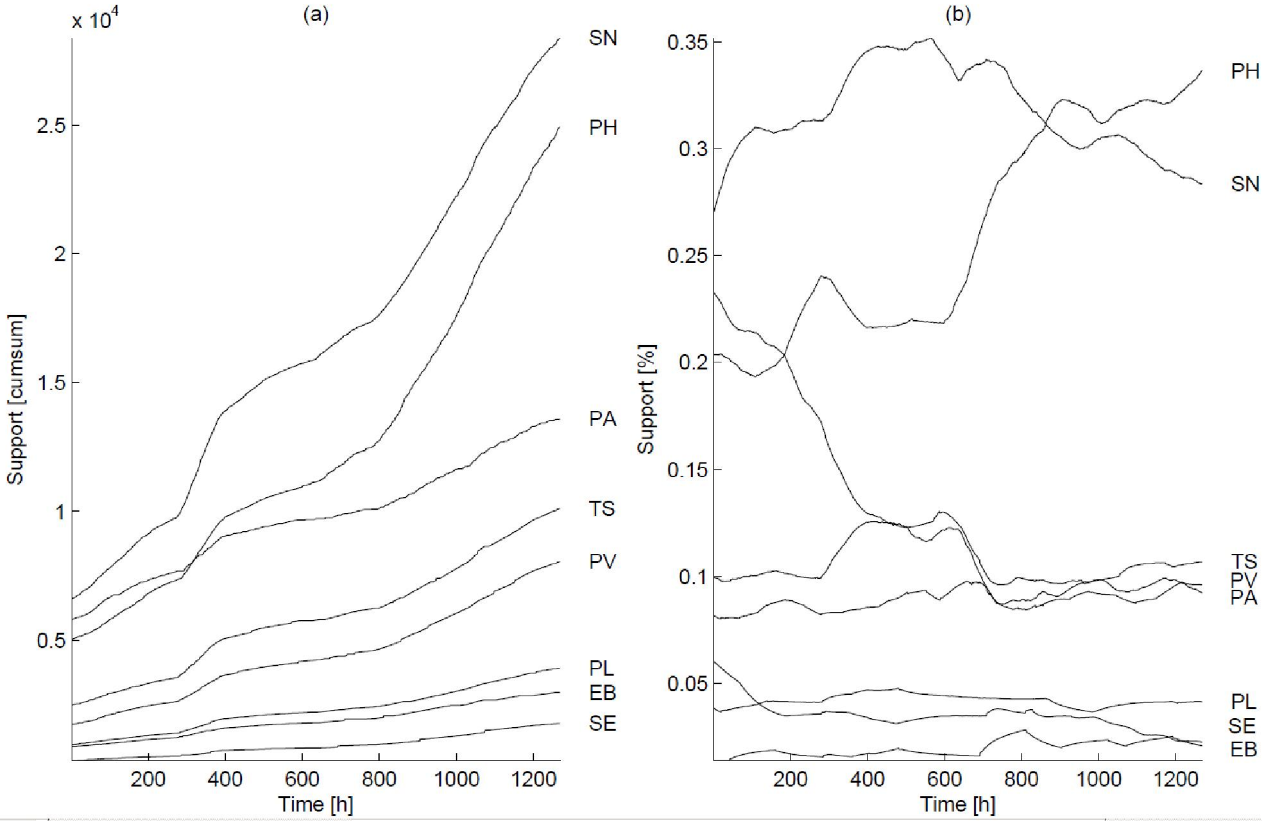


Fig. 4 – Support – will vote. (a) Cumulative support (matrix S). (b) Support matrix *V* values [hours]

6.1. COMBINING VAA AND GALLUP RESULTS

The difference of VAA users in time was calculated using values one day before *t* as *S*(*t*)–*S*(*t*–1), see (8). As it is seen in (9), we need to calculate *D*(*V*). In the experiments, the difference between three days in support was used *D*(*V*)=*V*(*t*)–*V*(*t*-3).

Approximated daily sample sizes were stored in matrix N^* (size 119 × 17), see (3).

Coefficient $\gamma(t)$ is a function of VAA and recursive Gallup sample sizes, see Fig. 5. With larger $\gamma(t)$ values the difference of VAA data is weighted more in the model *Z*, see (9).

Parameter values λ and α were used and weighted sample sizes for each Gallup in period *t*=1,...,119 were calculated, see (5) and Fig. 6.

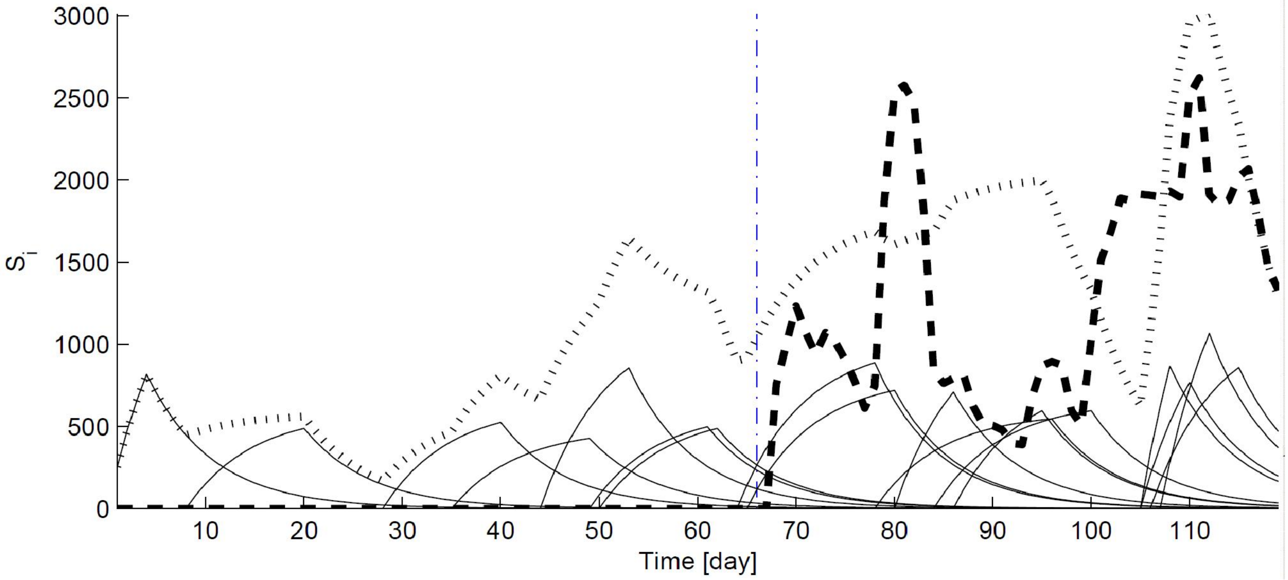


Fig. 5 – Recursive and cumulative Gallup sample size (dotted line), *D*(*S*(*t*)) derivate of VAA sample size (dashed line) and *N(*j*,*t*) recursive Gallup *j* sample sizes (solid lines)**

If newer polls are expected to be more important, λ should be smaller.Weighted sample sizes N^* are shown in Fig. 6 with optimized parameter values.

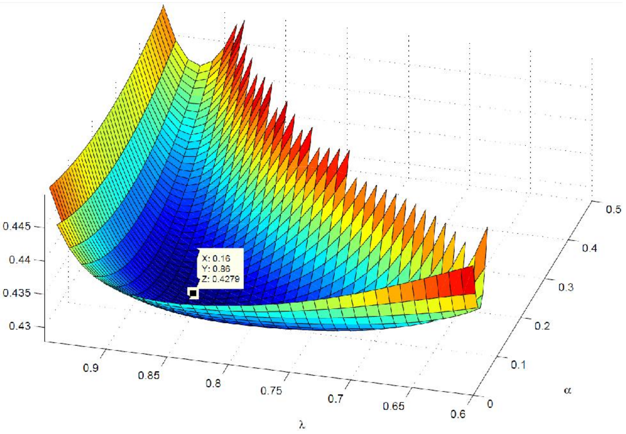


Fig. 6 – Optimization. MSE minimum is achieved when parameters are λ = 0.86 and α = 0.16

The non-value before the election date was large (25%), so forecasting the results is not meaningful. Nevertheless, we can mention that the order of the results was almost the same as the last row of matrix *Z*(119,:).

However, it is possible to examine news during the campaign and try to find explanations for changes in candidates' supports. Some news topics on Helsingin Sanomat are mentioned here. E.g., December 2, 2011 (*t*=68) Pekka Haavisto and the Finns Member of the Parliament Teuvo Hakkarainen had a media meeting and afterwards this was mentioned as good campaigning for Haavisto. On 9th of December (*t*=74) it was mentioned in news that the presidential candidates have large differences in the number of public Facebook supporters. During the campaign, Pekka Haavisto was the most popular candidate after Sauli Niinistö. Haavisto's "Obama Effect" was mentioned in the media and it seems that he mobilized young people to vote [34].

December 30, 2011 (*t*=95) Paavo Väyrynen started a TV ad campaign. It seems that the largest changes in support occurred on January 9-12, see Fig. 8. On the 9th of January (*t*=105) there was an audience rush in a panel discussion in Tampere and few days later a smear campaign started.

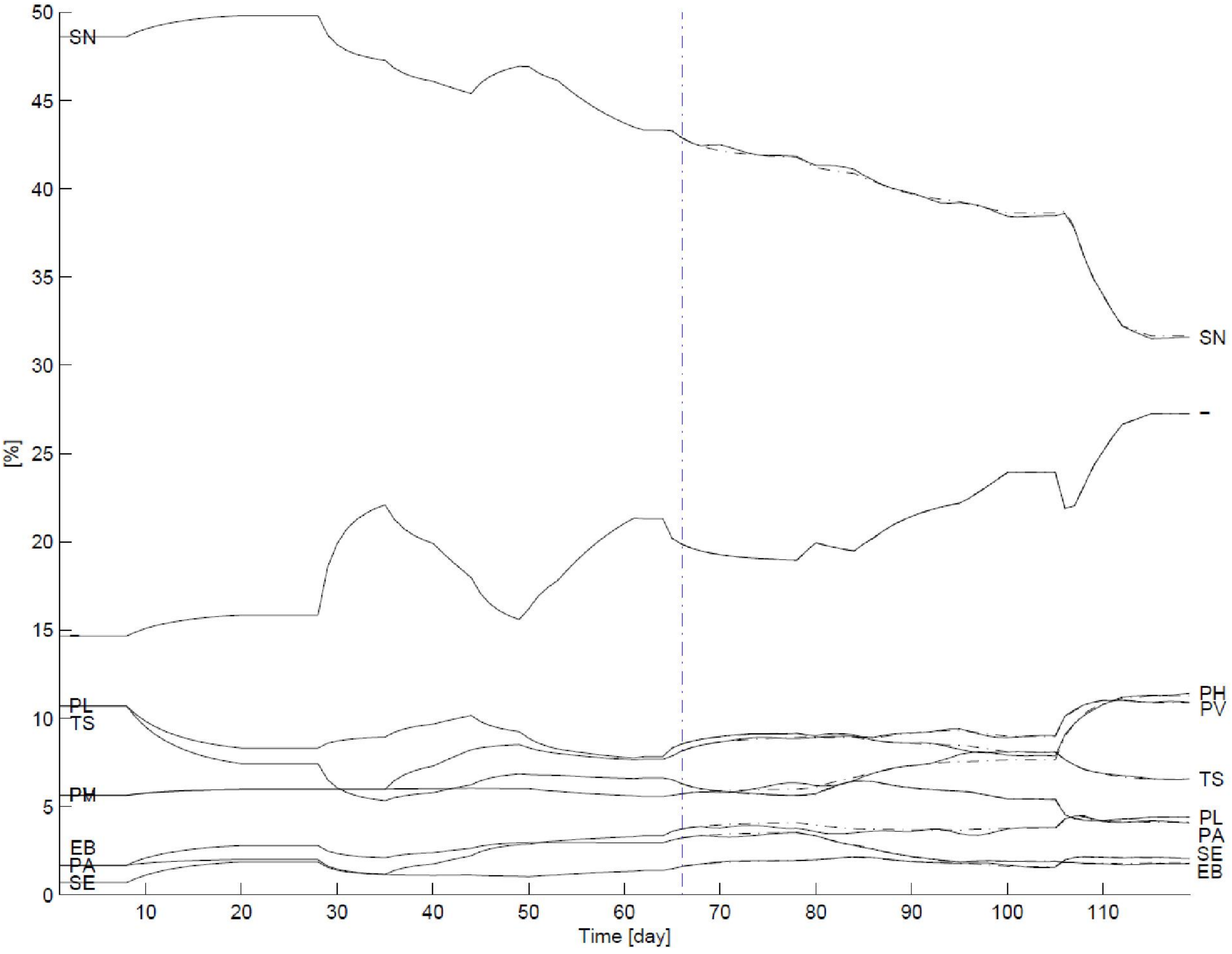


Fig. 7 – Gallup results and "will vote" probability combined. Gallup sample sizes and total samples (bold line) in time. Combined support of each candidate (solid lines) and support without VAA data (dashed line). Data is visualized from the starting date of first Gallup 26.9.2011 (*t*=1) until the election date 22.1.2012 (*t*=119). The dashed vertical line is the beginning of December (*t*=66), the first date when VAA data was combined to analysis

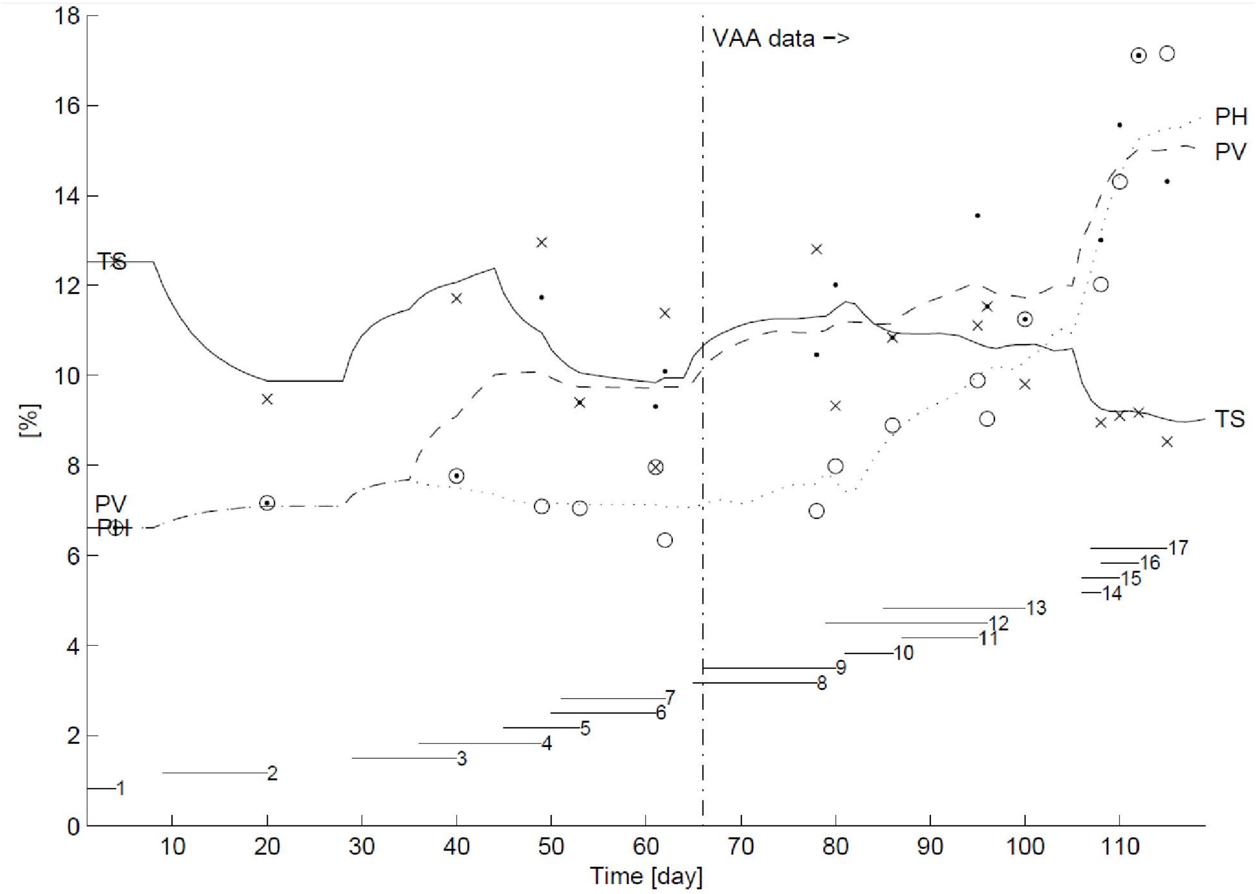


Fig. 8 – Three candidates were fighting for second position in the first round of election. The daily and Gallup support values for candidates are calculated using Gallup sample sizes $S_i\backslash non$. Timo Soini (solid line & x) was leading in the beginning. Paavo Väyrynen (dashed line & dots) and Pekka Haavisto (dotted line & circles) had good presidential campaigns. The horizontal numbered lines represents Gallups and the line length represents the duration of the poll

7. CONCLUSIONS

We introduced novel methods in this paper for combining temporal political support data sets and improved the candidates' support estimates with the proposed preprocessing methods. We paid attention to find the best parameters for a merged recursive model. The method presented in this paper allowed us to visualize the daily support of each candidate before the election. The results can be used for further research such as forecasting the results and analyzing the success of presidential campaigns. We also gave some examples which might have affected the campaigns. However, final conclusions about the campaigns are left for the readers and political experts.

8. DISCUSSION

In future, we plan to add more data sources to make it possible to do forecasts and comparisons between the results. Different poll sources could be analyzed.

E.g., by using parameter optimization different weights could be given to poll providers. In this work, all poll providers had equal weights: $W_{RIF}=W_{TNS}=W_{TT}=W_{MC}$. We also plan to study why the proposed (BM) candidates vary so much in time. This analysis was not included in this work since there was no information about the basic principles of the BM algorithm.

However, it will be hard to find out how voters have been maneuvering between the two rounds of the elections. Many conditions affect the results. E.g., how did different weather conditions between southern and northern Finland affect the voting activity? We do not give opinions about publishing Gallup results before election. In general, it can be said that after every election there is discussion about the effect of Gallups to the voting decisions – related to questions like: does it make sense to vote the candidate with fourth highest support.

9. ACKNOWLEDGEMENTS

The work was done in the AIRC Finnish Center of Excellence at the department of Computer and Information Science of Aalto University. Special thanks also for the graduate school HecSe and Helsingin Sanomat.

10. REFERENCES

[1] F. Hirzalla, L. Van Zoonen, and J. De Ridder, "Internet use and political participation: Reflections on the mobilization/normalization controversy," *The Information Society*, vol. 27, issue 1, pp. 1-15, 2010.

[2] A. J. Berinsky, "American public opinion in the 1930s and 1940s," *Public Opinion Quarterly*, vol. 70, issue 4, pp. 499-529, 2006.

[3] A. J. Scott and T. M. F. Smith, "Analysis of repeated surveys using time series methods," *Journal of the American Statistical Association*, vol. 69, issue 347, pp. 674-678, 1974.

[4] A. J. Scott, T. M. F. Smith, and R. G. Jones, "The application of time series methods to the analysis of repeated surveys," *International Statistical Review / Revue Internationale de Statistique*, vol. 45, issue 1, pp. 13-28, 1977.

[5] D. B. N. Silva and T. M. F. Smith, "Modelling compositional time series from repeated surveys," *Survey Methodology*, vol. 27, issue 2, pp. 205-215, 2001.

[6] D. Steel and C. McLaren, *Design and Analysis of Repeated Surveys*, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 11-08, 2008, 13 p.

[7] J. Talonen, M. Sirola, M. Sulkava, "Data fusion of pre-election Gallups and polls for improved support estimates," in *Proceedings of the 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS'2015*, Warsaw, Poland, September 24-26, 2015, vol. 1, pp. 845-849.

[8] G. Tsebelis, "Decision making in political systems: veto players in presidentialism, multicameralism and multipartyism," *British Journal of Political Science*, vol. 25, issue 3, pp. 289-325, 1995.

[9] R. A. Irvin, J. Stansbury, "Citizen participation in decision making: is it worth the effort," *Public Administration Review*, vol. 64, issue 1, pp. 55-65, 2004.

[10] F. Hirzalla, L. van Zoonen, J. de Ridder, "Internet use and political participation: reflections on the mobilization/normalization controversy," *Information Society*, vol. 27, issue 1, pp. 1-15, 2011.

[11] A. Graefe, S. J. Armstrong, "Predicting elections from the most important issue: A test of the take-the-best heuristic," *Journal of Behavioral Decision Making*, vol. 25, issue 1, pp. 41-48, 2012.

[12] S. Walgrave, M. Nuytemans, K. Pepermans, "Voting aid applications and the effect of statement selection," *West European Politics*, vol. 12, issue 6, pp. 1161-1180, 2009.

[13] S. Walgrave, P. van Aelst, M. Nuytemans, "The electoral effects of a popular vote advice application at the 2004 Belgian elections," *Actapolitica*, vol. 43, issue 1, pp. 50-70, 2008.

[14] R. E. Goodin, J. M. Rice, "Waking up in the poll booth," *Perspectives of Politics*, vol. 7, issue 4, pp. 901-910, 2009.

[15] R. D. Collings, L. G. Eaton, M. Hendrickson, "A meta-analysis of the 2004 campaign polls: An analogy to practice and publication in psychology," *Psychological Reports*, vol. 100, issue 3, pp. 847-856, 2007.

[16] A. J. Berinsky, "American public opinion in the 1930s and 1940s – The analysis of quota-controlled sample survey data," *Public Opinion Quarterly*, vol. 70, issue 4, pp. 499-529, 2006.

[17] R. R. Barr, "Populists, outsiders and anti-establishment politics," *Party Politics*, vol. 15, issue 1, pp. 29-48, 2009.

[18] A. M. Belchoir, "Explaning left-right party congruence across European party systems – a test of micro-, meso- , and macro-level models," *Comparative Political Studies*, vol. 46, issue 3, pp. 352-386, 2012.

[19] O. E. Danzell, "Political parties: when do they turn to terror?" *Journal of Conflict Resolution*, vol. 55, issue 1, pp. 85-105, 2010. doi: 10.1177/0022002710381065.

[20] S. Finn, E. Mustafaraj, "Learning to discover political activism in the twitterverse," *Technical Contribution*, vol. 27, issue 1, pp. 17-24, 2013.

[21] A. M. Belchoir, A. Freire, "Is party type relevant to an explanation of policy congruence? Catchall versus ideological parties in the Portuguese case," *International Political Science Review*, vol. 34, issue 3, pp. 273-288, 2013.

[22] K. Gemensis, "Estimating parties' policy positions through voting advice applications: some methodological considerations," *Acta Politica*, vol. 48, pp. 268-295, 2013.

[23] G. Jiglau, T. Burean, G. Badescu, "The ideological mapping of political parties in Romania. The relationships between dimensions of competition and ideological consistency," in *Proceedings of the ECPR General Conference*, Bordeaux, France, 2013.

[24] D. Schwarz, L. Schädel, A. Ladner, "Pre-election positions and voting behavior in parliament: consistency among Swiss MPs," *The Swiss Political Science Review*, vol. 16, issue 3, pp. 533-564, 2010. doi: 10.1002/j.1662-6370.2010.tb00440.x.

[25] The elections website of the ministry of justice, [Online]: http://www.vaalit.fi/, retrieved at December 2011.

[26] HS Voting Advice Application, [Online]: www4.vaalikone.fi/presidentti2012, retrieved at February 2015. (in Finnish)

[27] HS blog, [Online]: blogit.hs.fi/hsnext/ helsingin-sanomat-julkaisee-vaalikoneen-tiedot-avoimena-rajapintana, retrieved at February 2015. (in Finnish)

[28] Suomen presidentinvaalit 2012, [Online]: fi.wikipedia.org/wiki/Suomen_presidentinvaali_2012, retrieved at February 2015.

[29] Jari Pajunen, Presidentinvaalitutkimus 2012, [Online]: www.yle.fi/tvuutiset/uutiset/upics/ liitetiedostot/presidenttikyselytaulukot.pdf, retrieved at February 2015. (in Finnish)

[30] Generic ballot provides clues for 2010 vote, [Online]: www.Gallup.com/poll/124010/generic-ballot-provides-clues-2010-vote.aspx, retrieved at February 2015.

[31] Support visualization, [Online]: commons.wikimedia.org/wiki/File:Kannatus-graafi-1-06-01-2012b.jpg, retrieved at February 2015.

[32] L. Ljung, *System Identification Theory for the User*, PTR Prentice Hall, Upper Saddle River, NJ, 1999.

[33] E. Kreyzig, *Advanced Engineering Mathematics*, John Wiley and Sons, 1993.

[34] I. Volan, Look out for social media “Obama effect” in Finland’s presidential race, [Online]: socialmedianordic.wordpress.com/2012/01/29/look-out-for-social-media-obama-effect-in-finlands-presidential-race, retrieved at February 2015.

Miki Sirola *has been a Research Engineer in the Laboratory of Computer and Information Science in Helsinki University of Technology since 1998, currently the Department of Computer Science in Aalto University. He received MSc 1988 in Electrical Engineering (System Control and Automation) from Helsinki University of Technology, LicTech 1993 in Electrical Engineering (Automation) and DTech 1999 in Automation and System Technology (Automation) from the same university. Prior to Helsinki University of Technology, he worked at VTT (Technical Research Centre of Finland) Automation as research scientist (1987-1998) and at Institutt for Energiteknikk (OECD Halden Reactor Project) as research scientist (1992-1993). He was nominated a Docent in Computerized Decision Support in 2005 in the Department of Computer Science and Engineering in Helsinki University of Technology, and a Docent in the same field in Aalto University in 2011.*

Jaakko Talonen*received the degree of D.Sc. in Computer Science from Aalto University, Finland in 2015. His academic work related to machine learning, exploratory data analysis and visualizations in application areas such as industrial processes and political science lasted until Aug 2013. However, data science is passion for him and currently he makes career on industry. In addition, he allocates some of his free time for coding apps and exploring (open) data sets of his interests, e.g. food and beer.*

Mika Sulkava *received the degree of D.Sc. in Computer Science and Engineering from Helsinki University of Technology, Finland in 2008. He is a senior research scientist in Analyses of enterprises and markets in the natural resources sector group at Natural Resources Institute Finland.*

His main research interests are machine learning, data mining, and exploratory data analysis in the fields of economic and agricultural informatics.