What Voting Advice Applications can teach us about Voters and Elections

Popp, Raluca*[1], Horvath, Laszlo[1], Banducci, Susan[1], Coan, Travis[1] and Krouwel, Andre[2]

[1]University of Exeter, UK
[2]Vrije Universiteit Amsterdam

September 1, 2016

Paper to be presented at the 2016 ECPR General Conference in Prague, September 7 - 10, first draft, please, do not cite without permission

Abstract

VAA sites generate rich sources of data, potentially allowing researchers to address a range of questions related to voting, political attitudes and behaviour. However, the representativeness of VAA samples is questioned because they are opt-in rather than samples selected through randomly generated process. In this paper, we make a case for using VAA generated data to analyse voting intentions. We find that with proper statistical adjustment, non-representative samples can be used to forecast election outcomes, offering similar results to national representative samples. Our analysis proceeds as follows: first, using multilevel regression and post-stratification, we demonstrate how to use a voluntary sample to predict the results of a 'representative' sample. Second, to assess this technique, we employ two opt-in samples from UK 2015 GE and 2016 Brexit and two nationally representative samples, the post-election BES 2010 and BES 2015, compared against official election results and polls. We make a case in favour of requesting more demographic and political information at the forefront of the VAA, as it proves to be useful in forecasting election results.

Introduction

In the modern age, while the traditional modes of survey research have known an increasing disengagement trend, the use of opt-in online websites is flourishing. Voting Advice Applications (hereafter VAAs) are online tools that match

*R.Popp@exeter.ac.uk

voters to parties on the basis of issue positions. Through opt-in online samples, users share personal information that is easy to capture, allowing for analysis on public mood, political attitudes and election forecasting. The data collected through these websites lead to new ways of conducting social analysis that are more timely and cost-efficient than the standard government tracking polls.
Among the survey errors associated with web surveys, VAA generated data suffers greatly from lack of representativeness due to limited coverage and measurement error [Andreadis, 2014]. Alongside the self-selection bias into the VAA sample, there is an additional self-selection bias, because not all the VAA users proceed in filling out the extra-questionnaire. The double selection bias brings in respondents with higher levels of political interest, that cannot be compared with the population at large.
The lack of a systematic sampling method cautions the scholars of the VAA field to draw inferences from these data. However, the increased popularity of VAAs, together with the large data sets they generate, led to searching for solutions to the problem of representativeness. The approach we are testing in this paper is called multilevel regression and post-stratification (MRP). The idea behind MRP is to partition the data into thousands of post-stratification cells, to predict the voter intent at cell level using multilevel regression and lastly, to aggregate the cell level predictions according to the population's demographic structure **?**.
There are two main issues we are trying to tackle in this paper. Firstly, we explore the use of probability samples for weighting non-representative polls. Assuming that the source of bias in non-representative samples is given only by demographics, such as age or gender, the use of population totals from the latest available census would eliminate the problem. However, the bias in these surveys is often caused by non-demographic variables such as partisanship, which census data do not cover. Therefore, we need to rely on probability samples that provide these information. Secondly, rather than fitting a single model to a unique case, for the sake of a more general argument, we demonstrate model-based post-stratification on multiple cases and election contexts.
In this paper we show that using proper statistical adjustments, non-representative samples can be used for election forecasting, pairing the results from nationally representative samples. We proceed as follows: section 2 describes the data collected via *uk.electioncompass*, the VAA tool used in the UK 2015 General Election and *ukandeuref.electioncompass*, for the UK referendum on the EU membership. Both samples are highly biased towards young, highly educated and politically interested individuals, living in urban areas, for which we need to adjust the predictions. In section 3, we construct prediction of voter intent via multilevel regression and post-stratification. We conclude in section 4 by discussing the potential the VAA samples hold.

The *electioncompass* data

Data generated by VAA platforms can contribute to a better understanding of contemporary electoral behavior, at relatively low costs. The increasing number of VAA platforms and VAA users are providing large amount of data, granting access over users' political preferences and political parties' stances on policy issues.

Election Compass for the UK 2015 General Election was set by a team of academics from the University of Exeter and Swansea University, in collaboration with Kieskompas, the Dutch VAA provider. The tool was available a month prior to the election and it generated 25.000 users responses.

Besides political preferences, the VAA captured the users' party preferences through the propensity-to-vote (PTV) question, asked before the vote advice was displayed. The front page of the tool asked the users to report their age, gender, education and how sure they are they will vote in the upcoming general election, variables that are further used in the MRP procedure. Furthermore, 1200 users filled out the extra questionnaire, providing more detailed information on socio-demographic variables and political attitudes, such as interest in politics, party identification and past vote.

The second source of data comes from the Kieskompas' tool for the 2016 UK referendum on the EU membership. The VAA was launched 4 weeks before the referendum and attracted 25.000 visits from the UK, with 2.000 users filling out the extra questionnaire.

As the growing literature on VAA shows, there are several advantages these data hold. Firstly, the Internet is a medium that is uniquely well-suited for collecting politically relevant data. Data is collected more rapidly online than offline, and at lower costs. Collecting the data online also facilitates greater insights into temporal processes, namely collecting post election responses through follow up survey, that can be connected to in-campaign responses. Secondly, the VAA platforms facilitate collection of email addresses, that can be used for follow-up surveys, offering the possibility to link post-election to in-campaign attitudes. Lastly, VAA sites engage different groups within society, groups that are not accessible using conventional survey methods, such as minority and extremist party supporters, young people and minority groups, avoiding oversampling methodologies necessary to reach small groups that would rise the costs of the survey.

Despite these advantages, social scientists have been slow to realise the potential of VAA data as a source of dynamic public opinion data. The major stumbling block to the data is the self-selection bias VAA data holds as an opt-in sample, that can be overcome with statistical adjustments, such as MRP. Last but not least, the political information that is collected in addition to the demographic and geographic data proves to be extremely valuable when appropriate statistical procedure are employed, offering a new lens through which to look at VAA generated data.

Estimating voter intent with multilevel regression and post-stratification

At election times, dozens of opinion polls are conducted, in order to estimate election results. The polls are based on nationally representative samples, mostly using RDD with corrections for non-response, based on demographic characteristics, such as gender, age and education [Wang et al., 2015]. Here we describe two models developed for estimating opinions from national polls, using two non-probabilistic samples, the 2015 *uk.electioncompass* and 2016 *ukandeuref.electioncompass*. The MRP procedure requires two steps: first, fitting the model, then applying the model in order to estimate the opinions.

The first step requires fitting a regression model for the individual response y given demographics. The model estimates an average response θ_l for each cross-classification $_l$ of demographic and political variables. For estimating the GE 2015 voter intent, the model uses gender, age, education, political interest, party identification and vote in the 2010 General Election, thus $_l = 1, ..., L = 576$ categories. The model for the referendum follows the same structure of gender, age, education, political interest, party identification and vote recall from the 2015 GE, but it also includes the user's geographical location, as in the UK NUT regions. More about the reasoning behind including geographical data in estimating voting intentions in a subsequent section, where the Brexit model is explained in more detail.

Next, we post-stratify the VAA data to mimic a representative sample of likely voters. This weighting by population is called post-stratification. Is an analysis of an unstratified sample, and it requires breaking the data into strata, then reweighting as it would have been done if the survey was stratified. Whilst stratification is used to adjust for potential differences between the sample and the population using the survey design, post-stratification deals with such adjustments in the data analysis part [Gelman and Hill, 2006]. As mentioned above, we used BES 2010 and BES 2015 post election surveys (which are weighted with design weights, as well as post-stratification weights) as the gold standard for post-stratification. The estimated population average of the response y in any region $j$is

$$\theta_j = \sum_{l \in j} N_l \theta_l / \sum_{l \in j} N_l$$

with each summation over the post-stratification categories l. The core idea is to partition the population into cells based on combinations of demographic and political characteristics, then to use the sample for predicting the response for each category, and finally to aggregate the cell-level predictions up to a population-level prediction by weighting each cell by its proportion in the population [Wang et al., 2015]. The large number of categories is due to non-response adjustments that requires to introduce demographic information. Therefore, some of the categories will contain few or no data. However, if a multilevel model is fitted, this is not a problem [Gelman and Hill, 2006]. Each factor is automatically given a variance component. According to Gelman and Hill

[2006], this inferential procedure works well and outperforms standards survey estimates when predicting state-level (region-level, in our case) outcomes.
The simplest way to generate cell-level predictions is to average the sample responses within each cell. Assuming that within a cell, the sample is drawn at random, the predictions should be unbiased. However, as the number of cells increases, the cells become sparser and the empirical sample averages become unstable. This issue is addressed by generating cell-level predictions via a multilevel regression model. Known as multilevel regression and post-stratification, this combined strategy has been used to obtain accurate estimates for subgroups [Ghitza and Gelman, 2013, Wang et al., 2015]. Given that forecasting election results is a tradition that started in the US [Lewis-Beck and Rice, 1992, Rosenstone, 1983], most of the studies are focused on getting state-level opinions. The use of MRP in the European context is rather limited. MRP performs well in the US context, in estimating state-level results from a national poll of 1400 respondents [Lax and Phillips, 2009] and it can create estimates for states that are usually not surveyed [Kastellec et al., 2010]. Therefore our VAA samples for 2015GE and 2016 Brexit are expected to provide sufficient information for MRP, as well as the small samples, that we subset based on political variables.

Post-stratifying on demographic variables

The GE 2015 model

Applying MRP in this context requires two steps. First, we predict vote intent at the level of each post-stratification cell. The cells were generated by considering all possible combinations of gender (male and female), age (6 categories) education (2 categories, university and non-university), political interest (2 categories), party identification (6 categories) and vote in the 2010 General Election (6 categories), resulting in $_l = 1, ..., L = 1728$ categories for the model including the political variables, respectively 144 for the model including only the demographic information. We fit individual party vote shares as follows, using Gelman and Hill [2006] notation:

$$Pr(y_i) = y_i \in Cons|Cons., Lab., ..., Other) = \\ logit^{-1}(\alpha_0 + \beta^{age}_{j[i]} + \beta^{female}_{j[i]} + \beta^{edu}_{j[i]} + \epsilon_{ij})$$

The post-stratifying variables (age, gender and education) are included as random effects, so that we can generate predictions of $\hat{y}$ at the cell level by adding to the intercept the appropriate random effects at their appropriate levels.
The second step consists in multiplying the raw cell-level predictions by their population proportions, such as

$$\hat{y}^{PS}_{j=female-18-24y.o.-uniedu.} = \frac{N_j \hat{y}_j}{N_j}$$

and so on. The aim is to derive $\hat{y}^{PS}$ as

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where PS is the population share and the cell level probabilities are the inverse logit of the fixed intercept plus random effect combinations. Finally, to derive the party vote share prediction, we weight each cells' probabilities, and sum them up. Our point predictions are the median values of a 1000 simulated vote choice models per party. Compared with the official election results and polls, the results are the following:

Table1. Party vote share GE 2015

	large.VAA	GE2015	polls
Conservative	28.34	36.8	34.0
Labour	23.86	30.5	33.7
UKIP	17.86	12.7	12.6
LibDem	11.59	7.9	8.9
SNP	3.61	4.7	NA
Greens	11.8	3.8	4.8
Plaid Cymru	2.23	0.6	NA
N	9886		

Our predictions overestimate the vote shares of smaller parties, borrowing from the shares of the largest two parties; the most likely culprit, the lack of political post-stratifying variables, such as party identification, previous vote choice and political interest. While the role of partisan identification is intuitive from the voting behaviour literature Converse and Pierce [1986], political interest is a variable that explains self-selection bias in VAA generated data [Pianzola, 2014], therefore we should account for its effect. Our solution is to include these variables as random effects and also post-stratify on them, like in the model above. We develop this model in a subsequent section.

Given the particularities of the British electoral system, the translation of votes into seats is not clear cut. The 2015 GE provide a perfect example: while SNP needed just 25.970 to return a politician into parliament, UKIP needed 3.8 million votes, hitting a record on the vote/seat ratio [Nelson, 2015]. The vote shares are beyond the point in predicting electoral outcomes, particularly the shape of the government. Mostly, the outcomes are determined by electoral geography, meaning the distribution of party supporters across seats. This is the reason why SNP, with 4.7 vote share, but clustered in Scotland, got 56 seats in the House of Commons, while UKIP, dispersed across the UK, with 12.7 vote share got only 1 seat.

This points to the importance of accounting for region in our models. The *uk.electioncompass* does not contain geographical information beyond the country level. But in the analysis of the Brexit data, the NUT regions will be included as random effects and then used to create predictions at regional level.

The Brexit model

The Brexit polls offered varying predictions, with as much as 10% difference between Leave and Remain. With the actual results so close, it represents a real test for MRP and the quality of our prediction.
In predicting the vote intent for the referendum, we split the data into post-stratification cells just as before, creating all possible combinations of gender (2 categories), age (6 categories) education (2 categories), political interest (2 categories), party identification (6 categories), vote in the 2015 General Election (6 categories) and NUT regions (11 categories), resulting in $_l = 1, ..., L = 1728$ categories for the model including the political variables and 144 categories for the model with demographic and geographical information only. We predict the probability of voting "Leave" as follows:

$$Pr(y_i) = y_i \in Leave|Leave, Remain) =$$
$$logit^{-1}(\alpha_0 + \beta_{j[i]}^{age} + \beta_{j[i]}^{female} + \beta_{j[i]}^{edu} + \epsilon_{ij})$$

The logistic regression above gives the probability that any VAA user will vote "Leave" given the person's gender, age, education and NUT region. As in the previous model, our point predictions are the median values of a 1000 simulated vote choice models per party.

Table2. Share of Leave votes in the Brexit referendum

	large.VAA	Brexit
UK total	51.2	51.9
East Midlands	54.9	58.8
East of England	51.4	56.5
London	46.2	40.1
North East	54.2	58.0
North West	55.1	53.7
Scotland	43.5	38.0
South East	46.1	51.8
South West	50.0	52.6
Wales	57.1	52.5
West Midlands	57.3	59.3
Yorkshire and the Humber	50.6	57.7
N	3701	

The prediction of the Brexit referendum offers a very close image to the actual results, if we look at the national share. The NUT level predictions are relatively close, indicating the leaning towards Leave or Remain at the region's level, with the exception of the South East. Even so, the predictions are more closer to the actual results compared to the ones offered by the polls, where the Leave prediction varied between 39 and 55%, looking at the polls from June 2016.

Post-stratification on political variables

The GE 2015 model

Applying MRP to our VAA sample to get party vote shares for the 2015 GE gave predictions that are off by as much as 8%. As stressed above, the most likely culprit, the additional bias introduced by the users' high levels of political interest, as well the lack of control for party identification and previous vote choice, as vote choice predictors. As mentioned in the previous section, from the 25.000 visitors of the *uk.electioncompass* tool, only a small share filled out the extra questionnaire, providing information about previous vote choice, political interest and party identification (827 after discarding cases with missing data). We turn towards this sample to predict vote intent. Considering these variables, our model becomes:

$$Pr(y_i) = y_i \in Cons|Cons., Lab., ..., Other) =$$
$$logit^{-1}(\alpha_0 + \beta_{j[i]}^{age} + \beta_{j[i]}^{female} + \beta_{j[i]}^{edu} + \beta_{j[i]}^{pol.int} + \beta_{j[i]}^{pID} + \beta_{j[i]}^{voteGE2010} + \epsilon_{ij})$$

After introducing the political variables as random effects, we use them for post-stratification. *Figure1* presents the MRP vote share predictions from the 2015 GE model with demographic information only and the model including political variables, compared against the official election results and polls.

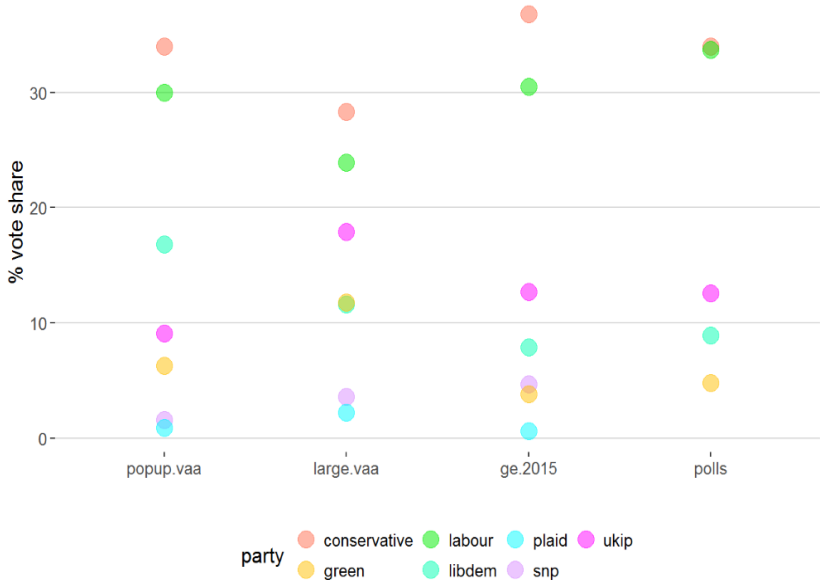

Figure 1: *Party vote share GE 2015*

Accounting for political interest, party identification and previous vote choice brings the predictions much closer to the election results (34.0% for Conservatives, 29.9% for Labour, 9.08% UKIP), but it still overestimates some of them (16.8% for LibDem). Although is does not mirror the official election results, the MRP with political variables gives more realistic predictions.

The Brexit model

Next, we extend the Brexit model to include political variables, and then create predictions at the NUT region level. Our model becomes:

$$Pr(y_i) = y_i \in Leave|Leave, Remain) =$$
$$logit^{-1}(\alpha_0 + \beta^{age}_{j[i]}) + \beta^{fem}_{j[i]}) + \beta^{edu}_{j[i]}) + \beta^{pol.int}_{j[i]} + \beta^{pID}_{j[i]} + \beta^{voteGE015}_{j[i]} + \epsilon_{ij}).$$

The second take on Brexit predictions, based on more refined post-stratification cells held worse predictions, with 35.7% national share for Leave. One possible explanation is the rather small number of cases. One the other hand, partisanship may not be a good predictor of the Leave vote. While UKIP took a clear stance on supporting Leave, both Conservative and Labour parties joined the Remain side. Given the Leave victory despite the support for Remain of the major parties, it indicates that partisanship may not have the same value in the referendum, as in the general election.
Figure2 illustrates the Brexit predictions, as an aggregate, and as the Leave vote share at the level of NUT regions. Although the aggregate results are far off, for some of the regions (see London, East of England and South West) it gives better predictions than the first model, using only demographic information. Given that the VAA samples are highly biased, the vote predictions are remarkably good. One should keep in mind that elections forecasting should be not only accurate, but also timely. VAAs are implemented about a month prior to the elections, allowing for this kind of analysis to provide elections predictions at relatively low costs.

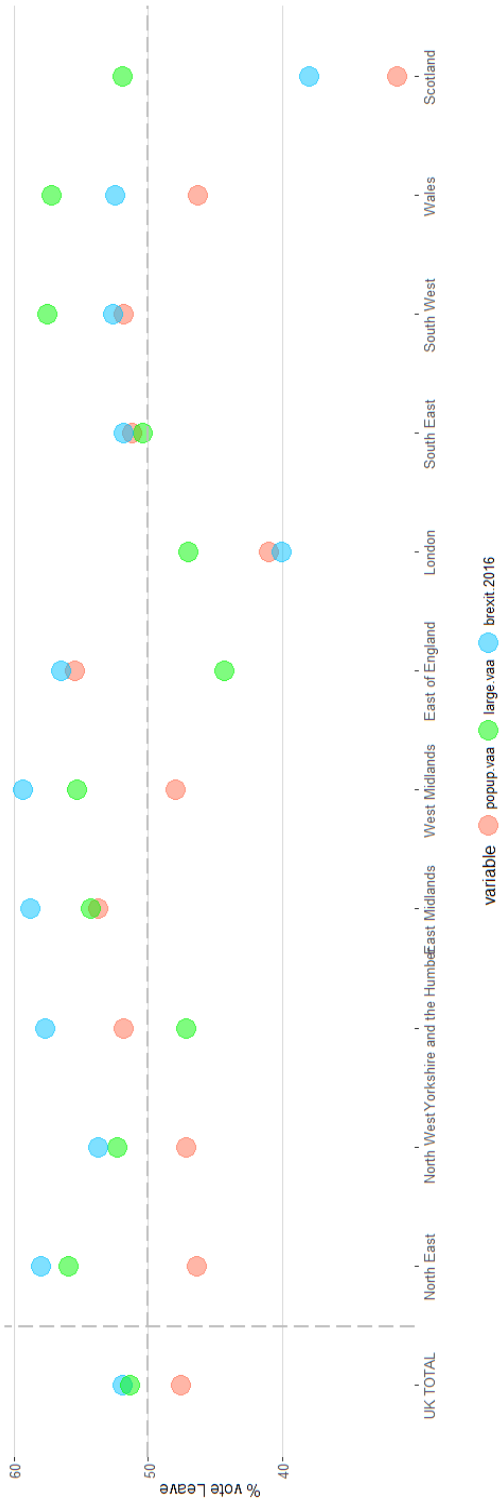

Figure 2: *Vote share "Leave"*

Conclusions

The aim of our paper was to show that with proper statistical adjustments, non-representative samples, such as VAA, can be used for election forecasting, matching the results from nationally representative samples. Using data collected through two VAA tools, *uk.electioncompass* for UK's GE 2015 and *ukandeuref.electioncompass* for 2016 Brexit, we constructed predictions of voter intent via MRP procedures.

Ideally, a non-representative sample would be post-stratified using census data, correcting for demographic mismatch. Unfortunately, in the case of VAA, the age, gender and education bias is overrun by another source of bias, namely political variables such as political interest and party identification. At this stage, the census data proves insufficient, so we must turn to nationally representative samples that include these variables. Given that these are weighted samples, the weights must be carried on in the subsequent procedures and accounted for. We stress the importance of data we rely on for post-stratification and use as a golden standard. Additionally, we show that including a variables such as geographical location has an incommensurable value. We made a point of testing MRP as a way of dealing with non-representative samples on more than one case. We argue this method improves inference from VAAs. Our major party vote predictions are much closer to the actual election outcomes than some of the official opinion polls, and our first referendum vote model yielded reasonable national vote share predictions. In our next steps we consider ways to derive prediction intervals through simulation and bootstrapping methods, as well as ways to include the BES sample error in the post-stratification process. Though we did use simulation to derive our point predictions, the variability of these values are relatively large reflecting overall model stability, rather than sensible prediction intervals. Thus we are still working on better ways to derive prediction intervals. Also, we will apply the same procedure to data from EUvox, a pan-European VAA for the 2014 European Parliamentary elections.

We conclude by stressing the importance of requesting demographic information at the front page of the VAA tool, without which correcting methods such as post-stratification would be impossible, and urging VAA developers to consider the potential benefits it would yield.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n 608085.

References

Ioannis Andreadis. Data quality and data cleaning. *Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective*, pages 79–93, 2014.

Philip E Converse and Roy Pierce. *Political representation in France*. Harvard University Press, 1986.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Yair Ghitza and Andrew Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

Jonathan P Kastellec, Jeffrey R Lax, and Justin Phillips. Estimating state public opinion with multi-level regression and poststratification using r. *Unpublished manuscript*, 2010.

Jeffrey R Lax and Justin H Phillips. How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1):107–121, 2009.

Michael S Lewis-Beck and Tom W Rice. *Forecasting elections*. 1992.

Fraser Nelson. *With 56 SNPs and just one Ukip MP, how can the Commons reflect the UK's political will?*, 2015. `http://blogs.spectator.co.uk/2015/05/with-56-snps-and-just-one-ukip-mp-how-can-the-commons-reflect-the-uks-political-will/` [Accessed: 23/08/2016].

Joëlle Pianzola. Selection biases in voting advice application research. *Electoral Studies*, 36:272–280, 2014.

Steven J Rosenstone. *Forecasting presidential elections*. Yale University Press, 1983.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.