Austrian Journal of Statistics
September 2018, Volume 47, 26–33.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i5.738

Handling Compositional Time Series with Varying Number of Parts

Jakob Bergman
Lund University

Abstract

When different polling organisations conduct political party preference polls at different times, different parties might be reported. If the estimated voter shares of these polls are combined into a time series we obtain a compositional time series, but with varying number of parts, thus prohibiting the use of standard compositional time series analysis tools. We discuss the problem and suggest a solution by imputing the unreported parts. The method is applied to a short compositional time series of party preference polls from Sweden.

Keywords: compositional loess, compositional time series, imputation, political party preference polls, polls Sweden.

1. Introduction

Political opinion polls play an important rôle in many countries, especially before an upcoming election or referendum. In a referendum or a bipartisan system it is usually clear what shares the polling organisations report, but in a multiparty system, especially with a multitude of small parties, the choice of which parties to report is at the discretion of the polling organisation (or the party commissioning the poll). This is can lead to different polling organisations reporting different parties, i.e. reporting vectors of proportions of different lengths.

A vector of proportions is a *composition*. We will not dwell on the theory of compositions (the interested reader is referred to Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015)), but merely note that a subset of a composition is called a *subcomposition*, that standard vector operations such as addition ($+$), scalar multiplication ($\cdot$), and norm ($\|\cdot\|$) have compositional counterparts, e.g. perturbation ($\oplus$), power transformation ($\odot$), and simplicial norm ($\|\cdot\|_S$), and that any composition or subcomposition $\boldsymbol{x}$ may be closed to equal any positive constant κ using the closure operation

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

Now returning to party preference polls, let the electorate be partitioned into D mutually exclusive parts, representing the different parties, and where one of the parts is an amalga-

mation of all the smallest parties, Other parties. Then the vector of shares for the D parts is the composition that polling organisations try to estimate. If this composition is measured repeatedly over time, we obtain a compositional time series. (For an overview of various models for compositional time series, see Aguilar Zuil, Barceló-Vidal, and Larrosa (2007).)

It has become common practice to combine some or all polls to attain better information of the political opinion, e.g. in order to predict the outcome of elections. However, if different polling organisations report different parties, or even the same organisation reports different parties at different time points, we end up with a compositional time series with varying number of parts. From a compositional point of view, a simple remedy would be to look at the subcomposition of parties, which are reported in all polls. However, if the goal of the analysis is a prediction of the election outcome, this is not really an option as we are trying to predict the full composition. Another option would be to amalgamate all partially reported parties with Other parties, but this is a less tractable option as we lose information. Furthermore, if the election threshold is low, small parties may be important to predict, since they can affect the distribution of seats in the parliament (assuming a proportional election system). For these reasons, we instead suggest imputing the unreported parties.

2. Compositions with varying number of parts

It might perhaps seem strange that the number of parts should vary in a compositional time series. There are however two main reasons for this. First, in many countries, if not all, the number of parties do vary over time; new parties are created, existing parties disappear, parties merge and split. In some countries this occurs rarely and in others more frequently. As an example of the latter, consider Denmark: during the eight general elections 1990–2015 only two elections were contested by exactly the same parties and only seven out of a total of 18 parties contested all the elections. If a party does not exist, it cannot be reported. This is comparable to a structural zero and requires explicit modelling. However, the proposed models for structural zeros (Aitchison and Kay 2003; Bacon-Shone 2003) are not applicable as they are designed for cross-sectional data. Zero imputation (see e.g. Palarea-Albaladejo and Martín-Fernández 2013; Martín-Fernández, Palarea-Albaladejo, and Olea 2011) could be a practical but formally questionable way of handling this type of varying number of parts.

Secondly, a party might simply not be reported. The choice of which parties to report is an arbitrary decision done by the polling organisation (or the party commissioning the poll). Depending of the electoral system, the choice could be more or less obvious. Very small parties that are very unlikely to win parliament seats or otherwise affect the election outcome are of course not very interesting to report. For instance, the Swedish Christian-Democratic party was founded in 1964 and received 1.5–2 % of the votes in the elections 1965–1982, but was not reported in the polls until the mid-1980's. This can most likely be explained by the fact that the election threshold for seats in the Swedish parliament is four per cent, and the Christian-Democrats were steadily clearly below that threshold. Polling organisations could possibly also be hesitant of reporting extremist or controversial parties in fear of being accused of (indirectly) supporting them.

3. Imputing unreported parts

We suggest that parties, that exist but are not reported in a poll, are imputed in order to obtain a compositional time series with the same number of parts for all measurements, thus enabling compositional time series modelling.

An important difference from zero imputation is that the missing parts are not zero but have been amalgamated with some part, usually Other parties. Hence, whereas zero imputation reduces all parts, only Other parties should be reduced when imputing unreported parts, otherwise the estimates will be biased. A second difference is that in zero imputations, the

missing value is usually replaced with a small constant value (possibly with added noise), whereas in our case the unreported parts may be assumed to change over time, requiring replacement values that are time dependent.

We propose that the imputation is done in the following way. Assume that there are n observations $\boldsymbol{y}_k$ at times t_k $(k = 1, \ldots, n)$ of which m observations $\boldsymbol{y}_i$ at times t_i contain one or more unreported parts. Which parts are unreported may differ between the m observations, however, all n observations are assumed to contain the part Other parties. The imputation consists of two steps: estimation and replacement.

1. *Estimation.* Use a suitable method to create an estimate (or prediction) $\hat{\boldsymbol{y}}_i$ of the entire composition for each time point t_i with an observation with unreported part(s). This estimation may be done in more or less sophisticated ways, e.g. by using the average (compositional centre) of the two surrounding complete measurements, by using some appropriate smoothing technique, or by applying some time series model.

2. *Replacement.* Replace the unreported part(s) and Other parties of $\boldsymbol{y}_i$ with the corresponding subcomposition of the estimate $\hat{\boldsymbol{y}}_i$ closed to equal the original Other parties of $\boldsymbol{y}_i$.

As an example, assume we have six parties (A, ..., F) and Other parties, but only four of the parties (A, ..., D) and Other parties are reported in the l^{th} poll, $\boldsymbol{y}_l = (y_A, y_B, y_C, y_D, y_{OP})$. We therefore predict the composition at time t_l, e.g. as the average of the two surrounding observations:

$$\hat{\boldsymbol{y}}_l = (0.5 \odot \boldsymbol{y}_{l-1}) \oplus (0.5 \odot \boldsymbol{y}_{l+1}).$$

The subcomposition of predicted values for parties E, F, and Other parties, $(\hat{y}_E, \hat{y}_F, \hat{y}_{OP})$, is closed to equal the observed Other parties y_{OP}. Finally, we replace Other parties of $\boldsymbol{y}_l$ with the subcomposition yielding the imputed composition

$$\left(y_A, y_B, y_C, y_D, \frac{y_{OP}\hat{y}_E}{\hat{y}_E + \hat{y}_F + \hat{y}_{OP}}, \frac{y_{OP}\hat{y}_F}{\hat{y}_E + \hat{y}_F + \hat{y}_{OP}}, \frac{y_{OP}\hat{y}_{OP}}{\hat{y}_E + \hat{y}_F + \hat{y}_{OP}}\right).$$

Depending on the choice of estimation method, the procedure can either be performed simultaneously for all m observations with unreported parts or one at a time. (In the latter case, the already imputed observations may be utilised for estimation at the risk of increased bias.) The exact choice of implementation of the procedure will thus necessarily be situation specific, depending on the data (degree of covariation, trend etc.), and the prevalence and pattern of unreported parts.

The proposed procedure ensures subcompositional coherence in the sense that all parts except Other parties and the imputed parts remain unchanged and that the relation between each unchanged part and the amalgamation of the imputed parts and Other parties is not altered.

4. A small simulation study

In order to assess how the proposed method works under different conditions, a small simulation study was undertaken. Four-part compositions were generated using three different models: a compositional simple linear regression model ($\boldsymbol{y}_t = \boldsymbol{\beta}_0 \oplus (t \odot \boldsymbol{\beta}_1) \oplus \boldsymbol{\epsilon}_t$), a compositional random walk ($\boldsymbol{y}_t = \boldsymbol{y}_{t-1} \oplus \boldsymbol{\epsilon}_t$), and independent observations ($\boldsymbol{y}_t = \boldsymbol{\epsilon}_t$). In all models, $\boldsymbol{\epsilon}_t$ were generated from a logistic normal distribution with zero mean and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.010 & -0.001 & 0 \\ -0.001 & 0.015 & 0.004 \\ 0 & 0.004 & 0.020 \end{bmatrix}.$$

As the temporal proximity of adjacent time points, and thus usually the similarity of the observations, also could affect the imputation, three different structures of time points were

Table 1: The mean and standard deviation (SD) of the mean absolute deviations for the 500 simulated imputations presented for each combination of time point structure, data structure, and estimation strategy.

Time point structure	Data structure	$\mathcal{C}$-loess estimation		Regression estimation	
		Mean	SD	Mean	SD
Consecutive	Regression	0.115	0.052	0.102	0.044
5/30	Regression	0.130	0.056	0.104	0.044
Random	Regression	0.123	0.057	0.100	0.043
Consecutive	Random walk	0.078	0.034	0.131	0.063
5/30	Random walk	0.206	0.094	0.452	0.245
Random	Random walk	0.206	0.123	0.449	0.261
Consecutive	Independent	0.124	0.053	0.107	0.046
5/30	Independent	0.138	0.060	0.109	0.046
Random	Independent	0.138	0.067	0.114	0.048

utilised: 25 consecutive time points (t = 101, ..., 125), five time points with a gap of five units followed by a gap of 30 units repeated five times (t = 1, 6, 11, 16, 21, 51, 56, ..., 221), and 25 randomly sampled time points between 1 and 250. One compositional time series was generated for each combination of data generating structure and time point structure and repeated 500 times, in total generating 4 500 time series. Part 3 and 4 in each time series were amalgamated for observation number 9, 19, and 21, i.e. two that in some cases could be quite close and one fairly distant from the others.

Two different estimation strategies were tried on all the time series: a compositional loess smoothing, $\mathcal{C}$-loess, (Bergman and Holmquist 2014) using the five temporally closest observations, and a compositional simple linear regression model using all 22 four-part observations. Figure 1 depicts the nine time series (before amalgamation) together with the imputed values for one of the 500 repetitions.

To quantify the deviation of the imputed values from the original values, we calculated the mean absolute deviation (MAD) as the sum of the simplicial norms of the differences between the original data and the imputed data MAD $= \sum_t \|\boldsymbol{y}_t \ominus \hat{\boldsymbol{y}}_t\|_S$, where $\hat{\boldsymbol{y}}_t$ is the imputed observation. This was done for each time series and imputation method; larger values imply a greater distance between the imputed and the original values. The mean and the standard deviation of the MAD of the 500 simulations are presented in Table 1 for each combination of time and data structure and for both estimation strategies. To get an idea of the magnitude of these mean deviations, one may e.g. note that the deviation between $(0.40, 0.30, 0.21, 0.09)$ and $(0.40, 0.30, 0.20, 0.10)$ is 0.113 and between $(0.25, 0.25, 0.27, 0.23)$ and $(0.25, 0.25, 0.25, 0.25)$ is 0.113. This would indicate that in many cases the imputed values are off by 1–2 percentage points, which if it were polls would most likely be acceptable. However, there are cases when the imputed values are off by ten percentage points, as seen in Figure 1. This is primarily the case when data are generated from a random walk but we do not have consecutive observations, hence making the observations hard to predict.

From Table 1 it would seem that neither of the strategies is outperforming the other. As might be suspected, a regression based strategy performs better than the $\mathcal{C}$-loess when the data are generated from a regression model, but when the data are generated according to a random walk the $\mathcal{C}$-loess is clearly the better strategy with much lower both mean and standard deviation. The two strategies are performing about same when the data are independent observations, the regression strategy doing slightly better than the $\mathcal{C}$-loess, especially when the observations are not consecutive. The last effect is most likely explained by the fact that the regression estimation was based on 22 observations compared to only five for the $\mathcal{C}$-loess.

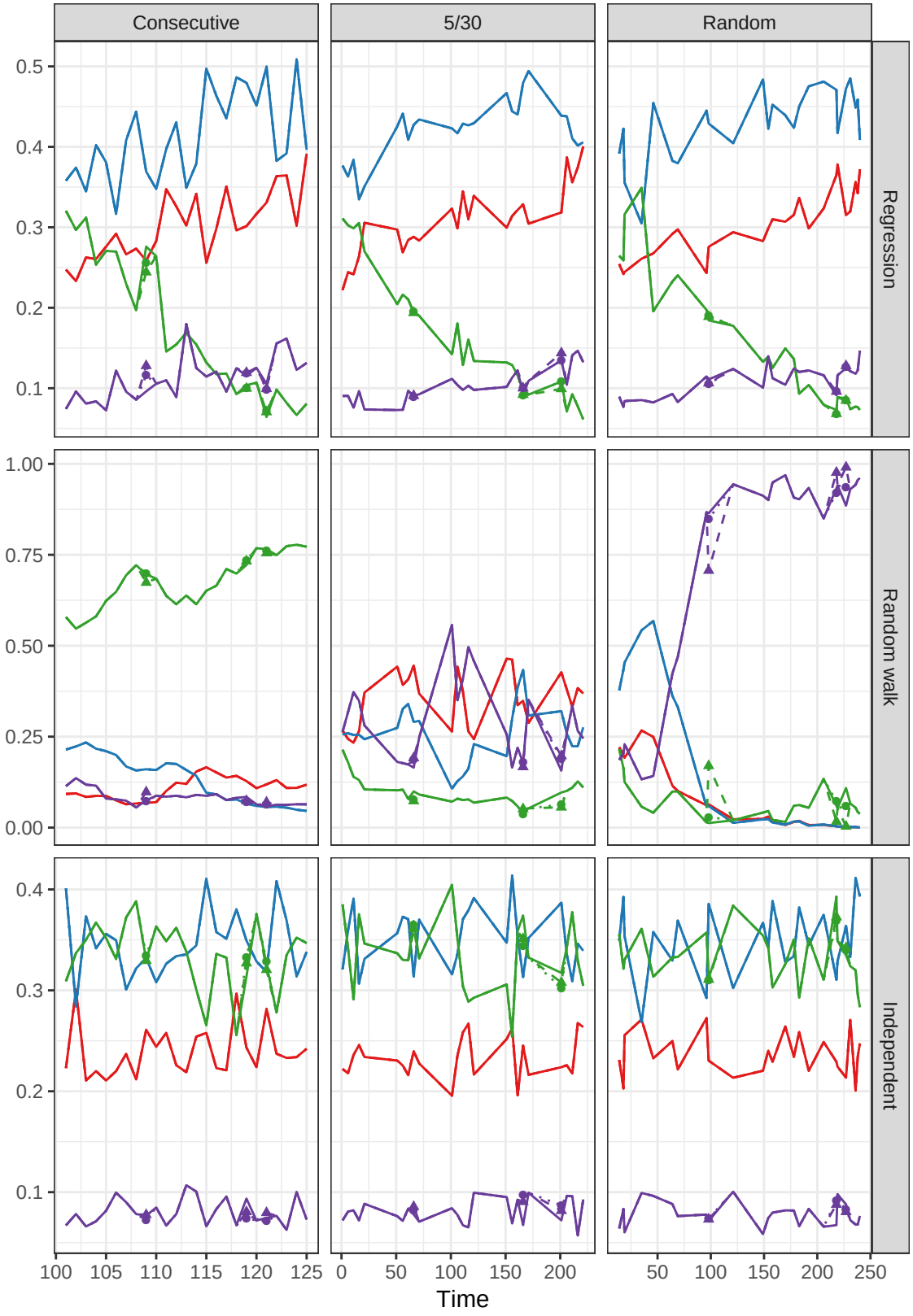

Figure 1: The original data (solid lines) and the imputed values using $\mathcal{C}$-loess (•) and using a regression model (▲). Part 3 is the green series and part 4 the purple series. Notice that in most cases the imputed values are quite close to the original. However, the middle right time series exhibits the largest observed error in the simulation; the regression strategy imputation being off by more than ten percentage points at the ninth time point.

5. An application

As an example we consider the 67 published polls in Sweden done during the six months prior to the general election on 14 September 2014. These include polls by eight commercial polling organisations, Statistics Sweden's party preference survey in May, and the exit poll done by the Swedish television. Data are available at www.novusgroup.se/vaeljaropinionen/ekotnovus-svensk-vaeljaropinion. The dates are the mid-date of the data collection period and not the publication date. In a few cases, the date has been altered one day earlier or later to avoid too much clustering. The time series is depicted in Figure 2. Notice the increased polling in the month before the election and the lack of polls during the summer.

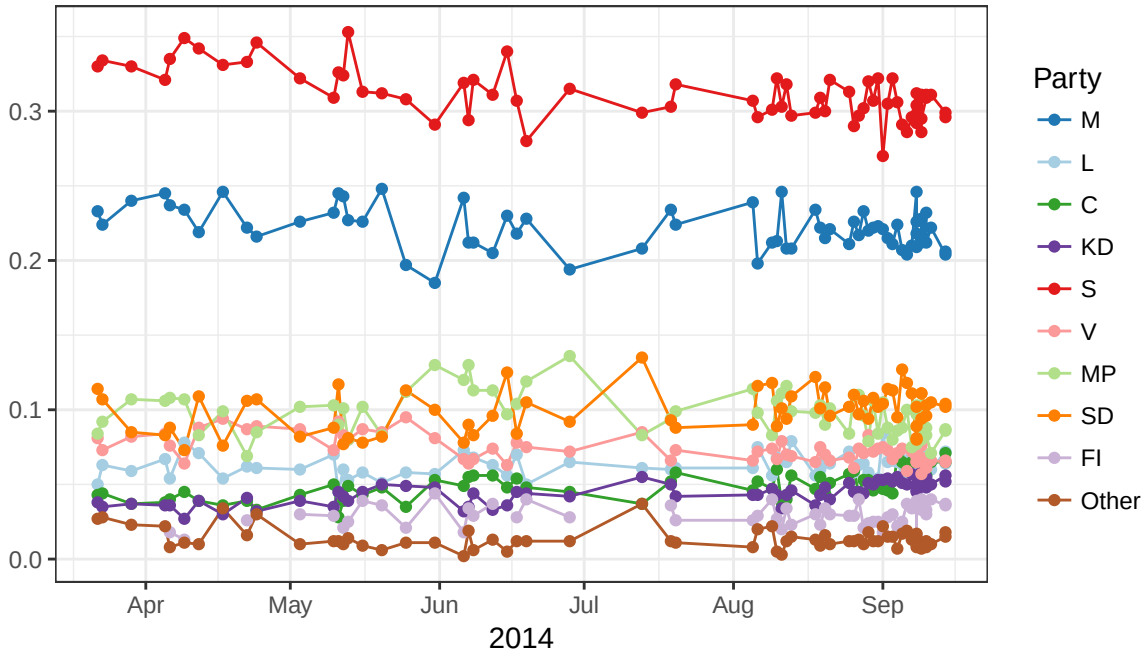

Figure 2: The graph shows the estimated voter shares for nine parties and Other parties from 67 polls published in Sweden from mid-March to mid-September 2014. Notice that the feminist party FI (light purple) is not reported in ten polls primarily during March and April, and that the amalgamation Other parties (brown) at these time points have a greater share than at the surrounding time points.

All the polls report the eight parties with seats in the Swedish parliament at the time. However, during the spring of 2014 the feminist party *Feministiskt initiativ* (FI) gained increasing support and began to be reported in the polls, with shares of 1.5–4 %. (The party did quite well in the European Parliament election in June 2014 with 5.5 % of the votes, winning one of the 19 seats. However, it should be noted that the Swedish voter turnout is generally much lower in elections to the European Parliament than to the Swedish Parliament.) Out of the 67 polls, 57 report FI. As can be seen in Figure 2, when FI is not reported the share of Other parties is much greater than for surrounding polls causing a clear shift of level in the graph.

The ten measurements where FI is not reported are predicted with $\mathcal{C}$-loess using the five temporally closest complete measurements. The subcomposition (FI, Other parties) of each prediction is then closed to equal Other parties of the respective original measurement, and Other parties is then replaced by the subcomposition. Figure 3 shows the result of the imputations. Apart from the obvious observation that FI now has measurements at all time points, one should notice that Other parties now lacks the level shifts, which seems more reasonable.

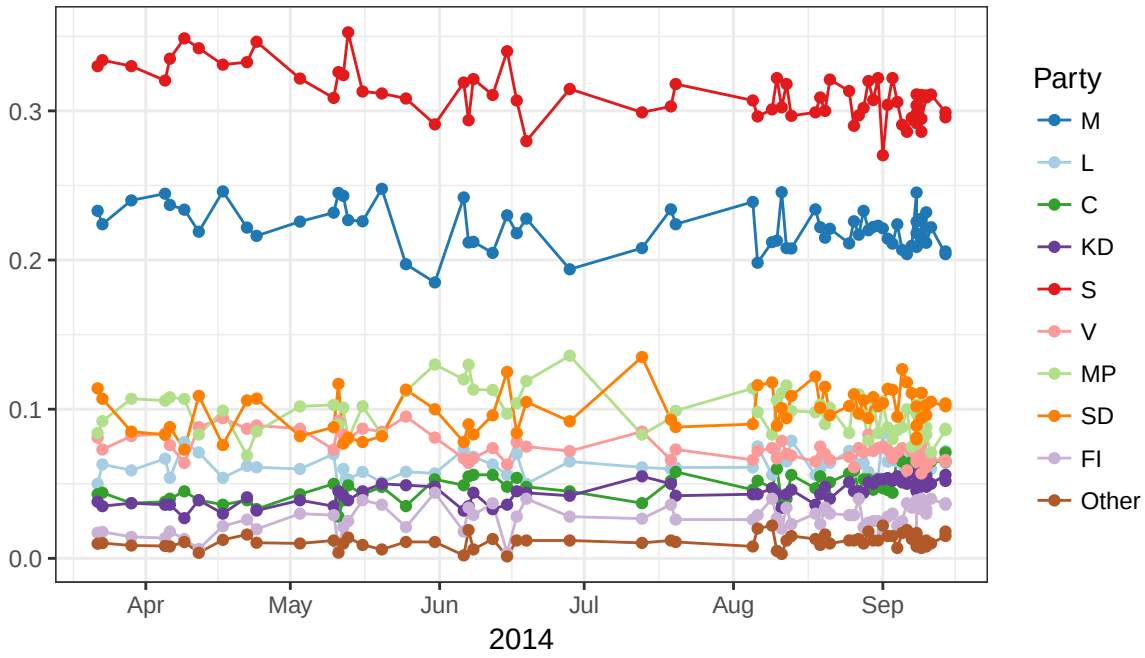

Figure 3: The same time series as in Figure 2 after the imputations. Notice that Other parties is now much smoother and that FI have measurements at all time points.

6. Summary

We have presented a method for handling compositional time series with varying number of parts by imputing the unreported parts. The method is developed for political party preference polls, but could of course also be used for any other compositional time series where analysing a subcomposition is not an appealing remedy for the problem of varying number of parts. We note that the usefulness of the method is highly dependent upon how the unreported parts are estimated; this in turn must be decided from case to case depending on the underlying data generating process.

It remains as future research to develop compositional time series models that can handle structural changes in number of parts, e.g. the creation of new, or mergers of existing, parties.

7. Acknowledgements

The simulations, applications, and illustrations in this paper were performed using R 3.4.3 (R Core Team 2017) and the packages composition (van den Boogaart, Tolosana, and Bren 2014), reshape2 (Wickham 2007), ggplot2 (Wickham 2009), and RColorBrewer (Neuwirth 2014).

The paper benefited from the comments and suggestions by two anonymous reviewers.

References

Aguilar Zuil L, Barceló-Vidal C, Larrosa JM (2007). "Compositional Time Series Analysis: A Review." In *Proceedings of the 56th Session of the ISI (ISI 2007)*. Lisboa, August 22–29.

Aitchison J, Kay JW (2003). "Possible Solutions of Some Essential Zero Problems in Compositional Data Analysis." In *Proceedings of CoDaWork'03, Compositional Data Analysis Workshop*. Universitat de Girona. URL http://hdl.handle.net/10256/652.

Bacon-Shone J (2003). "Modelling Structural Zeros in Compositional Data." In *Proceedings of CoDaWork'03, Compositional Data Analysis Workshop*. Universitat de Girona. URL http://hdl.handle.net/10256/661.

Bergman J, Holmquist B (2014). "Poll of Polls: A Compositional Loess Model." *Scandinavian Journal of Statistics. Theory and Applications*, **41**(2), 301–310. URL http://onlinelibrary.wiley.com/doi/10.1111/sjos.12023/.

Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011). "Dealing with Zeros." In V Pawlovsky-Glahn, A Buccianti (eds.), *Compositional Data Analysis. Theory and Applications*, pp. 43–58. Wiley.

Neuwirth E (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2, URL https://CRAN.R-project.org/package=RColorBrewer.

Palarea-Albaladejo J, Martín-Fernández JA (2013). "Values Below Detection Limit in Compositional Chemical Data." *Analytica Chimica Acta*, **764**, 32–43. URL http://www.sciencedirect.com/science/article/pii/S0003267012018363.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. Statistics in Practice. Wiley.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

van den Boogaart KG, Tolosana R, Bren M (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1, URL https://CRAN.R-project.org/package=compositions.

Wickham H (2007). "Reshaping Data with the reshape Package." *Journal of Statistical Software*, **21**(12), 1–20. URL http://www.jstatsoft.org/v21/i12/.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Affiliation:

Jakob Bergman
Department of Statistics
Lund University
Box 743
SE-220 07 Lund, Sweden
E-mail: jakob.bergman@stat.lu.se

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 47
September 2018

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2017-09-19
Accepted: 2018-04-24