A Syntethic Measurement for Political Engagement of Spending:

Pilot study to measure performance of local government using Open Government Data

Lorenzo Gabrielli*§, Davide Guido†, Fosca Giannotti* and Luca Bastiani‡
*National Council of Research - ISTI
Knowledge Discovery and Data Mining Laboratory, Pisa, Italy
Email: {name.surname}@isti.cnr.it
† University of Pavia
Department of Public Health, Pavia, Italy
Email: davide.guido@unipv.it
‡National Council of Research - IFC
Epidemiology Section, Pisa, Italy
Email: luca.bastiani@ifc.cnr.it
§University of Pisa
Department of Information Engineering, Pisa, Italy
Email: lorenzo.gabrielli@for.unipi.it

***Abstract*—Can we predict the outcome of future elections? Many politicians wonder if they will be re-elected and often they rely on polls conducted on a small sample of the population. In this work, we propose a data-driven approach that, given the past expenditure of city mayors, considers what are the most important aspects that determine the re-election. Our empirical results show the emergence of a particular expenditure threshold: mayors who invest in current and capital expenditure over such threshold during the mandate are more likely to be re-elected, while those who invest differently are more likely not to be re-elected. The impact of this research is to provide a new analytical tool that objectively shows to a public administrator if his actions will lead to the re-election.**

Keywords–Open Data; eDemocracy; eGovernment; Citizen-Government eModels, Data Mining, Administrative elections.

I. INTRODUCTION

The evaluation of the performance of public administration is a complex problem, for which the presence of Big Data and Open Data opens new scenarios. There are several works that, thanks to the digital traces left by all of us, studied the preferences of citizens in different economic and social areas [1] [2]. But, in the context of Public Administration, a new element to take into consideration is Open Data, in particular the data curated and made accessible by public administration. New data sources, in the field of political science, open up new scenarios for those who want to study the effectiveness of public administration in a data-driven approach.

As an alternative to the traditional surveys, several researchers have recently begun the use of social media such as Twitter to study the sentiment of voters before the election [3] [4]. On this aspect, other experiences led conflicting results, since users who use Twitter were not considered a representative sample [5]. The main limitation of this approach is that it is not applicable to small and medium realities. It is relatively easy to collect information on the voters' opinions about a national politician by means of social networks, but it is more difficult to do the same for all the mayors of each municipality.

In recent years, many governments began issuing the data as Open Data in order to ensure greater transparency of public administration. With this new data source, is possible to have the detailed expenditure of any public institution, in particular of Italian municipalities. In this paper, we propose a methodology that, using Open Data, produces a score and shows the trend of the expenditure managed by the re-elected and the not re-elected majors. To the best of our knowledge, there are no works that discuss this type of problem in a data-driven perspective, for administrative realities of small or medium size. This article is organized as follows: Section II describes the data used for the experiment, Section III introduces the analytical framework, Section IV is about the results obtained, Section V shows the main limitations of the approach and, Section VI details the impact of the research and the future works.

II. MATERIALS

The experimental dataset results from the processing of three different sources. (1) The dataset regarding the number of residents and size of Italian municipalities, that is provided by Italian Statistics Bureau (ISTAT) and is updated to the 2011 census. (2) The dataset on the expenditure of Italian local authorities that is provided by the Italian government [1]. Available expenditure of Italian municipalities for the years 2013, 2014 and 2015 (only first semester). (3) The dataset containing the election results of about 600 municipalities that voted in June 2015 plus the results of the previous consultations held in March 2010. The dataset is provided by the Ministry of Interior [2].

About the second dataset, the expenditure items are organized in a hierarchical manner, with the lower level having 248 items of expenditure. There are levels of intermediate aggregation respectively of 77, 34 and 4 families. In this phase of the study, the maximum aggregation level was used; the four families of expenditure (our endogenous variable) are (i) current expenses, (ii) cost of services for third parties, (iii)

[1] soldipubblici.gov.it
[2] storico.elezioni.interno.it

 ISBN: 978-1-61208-521-0

capital expenditures, (iv) expenses for loans repayment. (i) Current expenditure covers all public expenditure necessary to the ordinary activities of the state structure (eg. staff, purchase of consumer goods). (ii) Expenses for services concerning transactions carried out on behalf of third parties as the institution acts as withholding agent. (iii) The capital expenditures are the ones in which the State aims to play an active policy in the economy (eg. buying movable and immovable assets, shareholdings). (iv) The costs of repayment loans consists of repayments of loans and cash advances. In our model, all the values of the costs are expressed on a monthly average basis, and are scaled to the number of residents of each municipality. For space reasons we only show the distribution of value for the period before election (Figure 1). The distribution in the other years shows a similar trend.

The third dataset allows to understand if a mayor or a coalition government was reconfirmed. For about 350 municipalities only, it is possible to understand if the mayor (or party) is re-elected and only 70 of these have more than 15,000 inhabitants. There are two criteria for determining whether a mayor or a party won the election, namely: (i) checking the name of the former mayor and the new one or (ii) analyzing the political area of the previous and new mayor, inferred from the name of the political party that won the elections. As mentioned above, only for 350 municipalities out of 600 it is possible to infer if a mayor was re-elected as if (i) it was not valid, (ii) had indicated a general civil list. In the municipal elections, especially in small ones, it is common that all candidates use civic lists without party symbols. For this reason, the mayor of the municipality always belongs to a civil list, preventing to understand whether this list belongs to a party of the left, right or center wing. For these 350 municipalities, in 45% of cases, the mayor or the coalition is reconfirmed, while in the remaining 55%, it is substituted.

III. METHODS

The idea underlying our approach is that a single-factor model, synthetizing a set of families of expenditure, could facilitate the understanding of the expenses role to explain the political success.

Firstly, we used exploratory factor analysis (EFA) to investigate the expenditure factor structure [8]. Then, we used confirmatory factor analysis (CFA) to validate the factor structure provided by EFA. For each statistical unit (municipalities), we computed a synthetic score of"political engagement of spending" for each year using CFA [6].

Finally, were performed analysis of variance for repeated measures (ANOVA-Rm) for a single factor (time) stratified for the binary political success indicator, and t-tests to compare the "political engagement of spending" and the detached families of expenditure with a binary political success indicator (mayor re-elected or not) over time. At the end, for the year preceding the elections, we validated the "political engagement of spending" using Receiver Operating Characteristic (ROC) curve analysis. All statistical analyses were performed using R software (version 3.0.2 for Windows) [7].

A. Confirmatory Factor Analysis

A Confirmatory Factor Analysis (CFA) via Structural Equation Modeling (SEM) was performed to confirm the presence of a latent variable (factor) underlying the spending policies. The items (endogenous variables) are ($v1$) current expenses, ($v2$) cost of services for third parties, ($v3$) capital expenditures, ($v4$) expenses for loans repayment. The factor has been interpreted as "political engagement of spending". The coefficients linking the factor with items by linear equations are called "factor loadings".

The goodness of fit was evaluated with two indexes: (i) Standardized Root Mean Square Residual (SRMR); (ii) Root Mean Square Error of Approximation (RMSEA). In general, a model is considered to show good fit when the SRMR and RMSEA are not higher than 0.10. In addition, further CFA indexes as *gamma* for unidimensionality (acceptable if > 0.2) and the *general validity coefficient* (acceptable if > 0.8) verify the adequacy of single-factor measurement model. Finally, *the standardized factor loadings* (item-factor correlations) (acceptable if > 0.4) proved high-quality specific factor validity for each item.

Hence, the score validation (only for score of 2015) was performed by ROC analysis to evaluate the effectiveness of the (factor) score in distinguishing if the mayor was re-elected or not. Overall, the predictive performance was measured by the area under the curve (AUC): ROC curves with AUC of 0.5 indicating no predictive diagnosis versus AUC of 1.0 indicating perfect ones. Only the score of 2015 was validated because 2015 is the year before the elections, and expenses are directly related to the election campaign.

B. Analysis of variance and independent t-tests

The analysis of variance for repeated measures (ANOVA-Rm) is a useful multivariate method to evaluate changes of a continuous variable in relation to a categorical one, over the time. In this case, it was performed to assess the change of the score (obtained via CFA-SEM) over time, i.e., from 2013 to 2015, in relation to binary political success indicator (mayor confirmed or not).

Independent t-test is a useful statistical test to verify if the means of continuous variable across 2 independent groups are significantly different. It was performed to verify the differences of the mean of the score for each year, i.e., the differences of the families expenditure means between the 2 groups over the year.

IV. CASE STUDY: ITALIAN ADMINISTRATIVE ELECTIONS

The starting assumption is that it is possible to synthesize all expenditure in a single factor. The factor analysis (EFA) confirming the one-dimensionality of the concept underlying the costs taken into account, i.e. eigenvalue > 1 and Horn's parallel analysis identified one factor.

We apply the CFA-SEM model on the dataset regarding the administrative election held in Italy in June 2015. Our aim is to shows that families of expenditure contribute with different weights in the construction of the factor "political engagement of spending". Although standardized factor loadings of all the variables of the model are acceptable, the most important variable is the current expenditure, followed by capital expenditure, expenses for services for third parties and refund loans. This rank is the same for the three years.

Figure 2 shows the triples of the standardized factor loadings for the three years of analysis. The high-quality specific factors are those higher than 0.4. For example, regarding

 ISBN: 978-1-61208-521-0

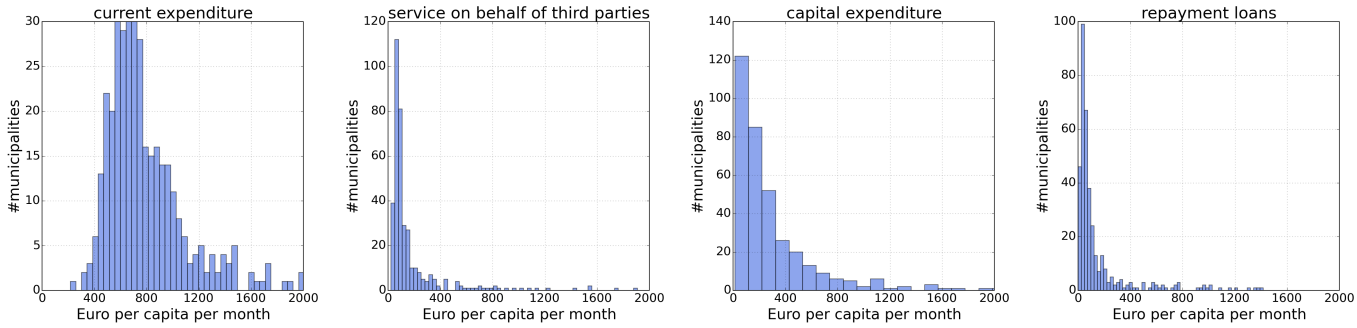

Figure 1. : Distribution of families expenditure per month per capita (2015)

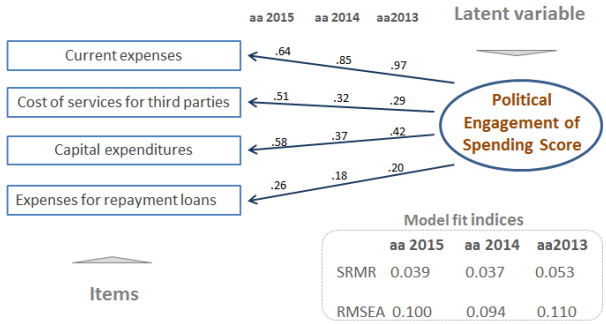

Figure 2. : Standardized factor loadings and goodness of fit indexes for 2013, 2014, 2015 models. The goodness of fit indexes (SMSR, RMSEA) of the "political engagement of spending" are acceptable for each year because they are less than 0.10

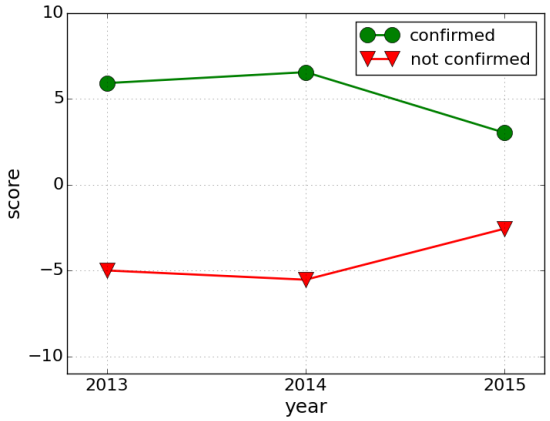

Figure 3. Plot of means of cPE score over 3 years. The relevance of this plot is confirmed by the ANOVA test that observes that within each group trends of the measure cPE are very similar. Instead comparing the cPE in each year between the two groups (green vs. red) with independent t-student, the differences are significantly different for each year (2013 $p = 0.019$; 2014 $p = 0.012$; 2015 $p = 0.007$).

current expenditure (the most important variable), values are respectively 0.97, 0.85, 0.64 for the years 2013, 2014, 2015. The goodness of fit (SMSR, RMSEA) of the "political engagement of spending" for each year is acceptable because it is lower of 0.10. Finally, we compute CFA indexes to measure internal validity of the model; *gamma* is higher than 0.2 (0.33) and *general validity coefficient* is higher than 0.8 (0.8). Thereby, we obtained a validated synthetic continuous score of "political engagement of spending" (centered factor score) (cPE) that we analyzed.

Figure 3 shows the trend of the means of score stratified for the binary political success indicator, divided into mayor confirmed or not confirmed. With the aim of assessing the trend of spending for mayors confirmed and unconfirmed, the analysis of variance (ANOVA) for repeated measures was performed. ANOVA did not show significant differences ($p > 0.05$) of cPE within the trend of the same group (2013 vs. 2014, 2014 vs. 2015). Comparing instead the cPE (in each year) between the two groups with independent t-tests (adjusted by Bonferroni method for multiple comparisons), the differences are significant for each couple of years (2013 vs. 2013, 2014 vs. 2014, 2015 vs. 2015).

Finally, Figure 4 shows the evaluation of individual families of expenditure during the three years with independent t-tests. In the trend of three years, confirmed mayors spent more in current expenditure ($p < 0.05$) and, in 2015, they also invest more in capital expenditures ($p < 0.05$).

Thanks to parameter estimates extracted from the CFA-SEM model 2015 (factor loadings and residual variance) an equation was defined (1) to compute the (rescaled) "political engagement of spending" score (rPE) of each municipality:

$$rPE = 0.0276 * v1 + 0.0139 * v2 + 0.0078 * v3 + 0.0155 * v4 \quad (1)$$

Then, to validate rPE and determine its cutoff to predict the election result, ROC curve analysis was performed. The optimal cutoff for rPE is 2.19, the AUC was 0.62 indicating an acceptable predictive accuracy. The sensibility was 0.649 and the specificity was 0.591. It is worth to remark this is just a preliminary result and further validation steps are needed, such as cross validation with one-leave-out methodologies and/or stratifying the datasets over a different dimension/wealth of Local Governments.

The remarkable result of our work is that city administration can use per capita expenditures sustained in the period before the elections, in order to understand, using (1), if the investment is similar of the mayors who were re-elected in the past.

 ISBN: 978-1-61208-521-0

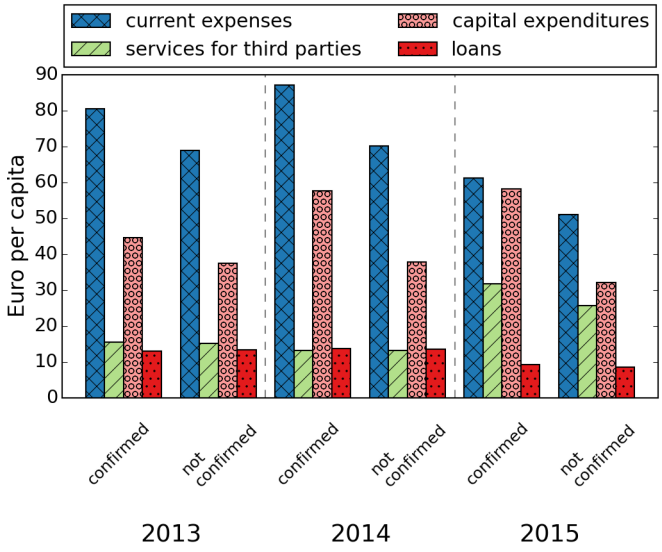

Figure 4. : Evaluation of individual families of expenditure during the three years with independent t-student. In the trend of three years, confirmed mayors spent more in current expenditure (2013 $p = 0.02$; 2014 $p = 0.010$; 2015 $p = 0.013$) and in 2015 also invest more in capital expenditures ($p = 0.025$).

V. Limitations

The main limitation of this approach is the difficulty of collecting the election results of the previous years. This limitation is overcome by the fact that, starting from 2014 the Ministry of Interior has provided an open data service, so, hopefully, in the future there will be more data available. In addition, the model can be specialized considering the political party affiliation of the mayor. At the moment, the Open Data website [3] does not provide the income of the municipalities, i.e., the taxes paid by citizens; with this information, the model could be further enriched.

VI. Learned lesson and next steps

Our work shows how Open Data of the public administration can be used to anticipate the judgment that citizens will express in the next election. The contributions of this study are: (i) provided different weights for the families of expenditure through the CFA-SEM model, (ii) build a synthetic measurement of expenditure called "political engagement of spending" score, validated using the election results, (iii) estimate a score cutoff to predict if an administration will be, more probably, confirmed or not.

The CFA-SEM model shows different weights (standardized factor loadings) among items of expenditure for the administrations confirmed or not confirmed. In the six months before the election the weight of current expenditure and capital expenditure is higher in municipalities that will be confirmed. This trend was detected in relation to binary political success indicator both overall, considering the score of "political engagement of spending" (Figure 3), and by comparing the detach averages of the municipal expenses in the various years. In addition, our framework provides an indication of the critical thresholds of individual cost items, on which one needs to take action in order to improve the overall score (Figure 4).

This result is of great impact, because a single administrator, by entering into a formula (1) expenditure per capita in four families can measure the "likelihood" of his reelection. Thinking in monetary terms, the threshold values that indicate a higher likelihood of re-election could be about $100Euro$ per month per capita. This suggests that does not pay to invest only in the last six months. The rank of expenditure, per capita per month, is: current expense, i.e., the provision of main services to citizens ($45Euro$), the second is cost of services for third parties ($27Euro$), $12Euro$ for expenditures capital and $6Euro$ for repayment loans. Details will be further investigated with more complete data.

In the future, we plan to repeat the analysis by considering a greater level of detail of spending that compose each of the four families utilized. In addition, the experimental dataset will be enriched with the morphological and productive datasets (wealth, well-being, etc.) of the area, the seasonal trends of expenditure and income, the political party and subjective variables, such as the self-perception of the citizen on the effectiveness and efficiency of their own mayor.

[3] soldipubblici.gov.it

VII. Acknowledgments

Thanks to the technicians of the Ministry of Interior for providing data of the past elections. Thanks to Pierpaolo Galleni, Fabio Coppede' and Francesca Pratesi for their hints in the preparation of work. This work is partially supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem'. (http://www.sobigdata.eu).

References

[1] Pennacchioli D, Coscia M, Rinzivillo S, Giannotti F, Pedreschi D. The retail market as a complex system. EPJ Data Science. 2014

[2] Rinzivillo S., Gabrielli L., Nanni M., Pappalardo L., Pedreschi D. and Giannotti F. The Purpose of Motion: Learning Activities from Individual Mobility Networks. In: DSAA 2014

[3] Bermingham, Adam and Smeaton, Alan F. (2011) On using Twitter to monitor political sentiment and predict election results. In: Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), 13th November 2011, Chiang Mai, Thailand..

[4] Murphy Choy, Michelle L. F. Cheong, Ma Nang Laik, Koo Ping Shung: US Presidential Election 2012 Prediction using Census Corrected Twitter Model. CoRR abs/1211.0938

[5] Alexander Furnas . You Can't Use Twitter to Predict Election Results. The Atlantic. 2012

[6] Kline RB. Principles and practice of structural equation modeling, vol. 156; 2011.

[7] Development Core Team (2012) A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna.

[8] Horn JL. A rationale and a test for the number of factors in factor analysis. Psychometrika 1965;30:179185.

 ISBN: 978-1-61208-521-0