ITM Web of Conferences **6**, 02002 (2016)
DOI: 10.1051/itmconf/20160602002

Cluster analysis of the bias in the SMB subsidy recipients selection

Igor Balk[1, a] and Egor Mateshuk[2]

[1]*Global Innovation Labs, USA*
[2]*Moscow Institute of Physics and Technology, Moscow, Russia*

Abstract. This paper is discussing the applicability of cluster analysis techniques to studying bias in the selection process of government SMB subsidy recipients using data from Moscow city SMB administration. Ward's Hierarchical Agglomerative Clustering Method is used to study 812 SMBs, which received subsidies in 2012-2013 from the Moscow City SMB administration. This analysis strongly demonstrates that clustering mechanisms can be used to reduce selection irregularities in government subsidy distribution.

1 Introduction

Governments around the world are making a significant effort towards stimulating various initiatives aiming to boost economic development of their countries. Naturally, this raises the question if taxpayer's money spent on such programs is achieving desired goals. Therefore, the question of effectiveness of various government projects around the world, which are designed to boost the economy, is widely discussed in the literature. One of the ways to measure this is to study treatment effects of various programs on boosting innovation and economic development as suggested by Czarnitzki [1], Hahn [2], Lechner [3] and others. The main problem when dealing with this approach is valid estimation of selection bias and correct selection of control groups. Balk [4] suggested using cluster analysis methods to properly select control groups, which provide the closest match to the group which received the treatment. In this paper we will concentrate on discussing the application of clustering algorithms to study selection bias for treatment of small and medium size businesses by Moscow Small Business Administration.

2 Overview of the clustering algorithm

In this study we have used Ward's Hierarchical Agglomerative Clustering Method, which was first introduced by Ward [5]. He suggested an agglomerative hierarchical clustering algorithm, where the criterion for choosing the pair of clusters to merge at each level is defined by the optimal value of an objective function.
We define a distance as a positive, definite, symmetric mapping of a pair of observation vectors onto the positive real, which in addition satisfies the triangular inequality. For observations *i, j, k* we have:

$$d(i,j) > 0, \quad (1)$$
$$d(i,j) = 0 \Leftrightarrow i = j, \quad (2)$$
$$d(i,j) = d(j,i). \quad (3)$$

For an observation set, I, with $i, j, k \in I$ we can write the distance as a mapping from the Cartesian product of the observation set into the positive real:

$$d: I \times I \to \mathbb{R}^+. \quad (4)$$

At each step of the clustering algorithm we find a pair of clusters that leads to a minimum increase in total within cluster variance after the merging, i.e. weighted squared distance between cluster centers

$$\frac{1}{|q|}\sum_{i\in q} d^2(i,q^*), \quad (5)$$

where q denote the cluster (a set) and q^* the cluster's center, which is defined as

$$q^* = \frac{1}{|q|}\sum_{i\in q} i. \quad (6)$$

Initially all clusters contain only a singe point. We can define Euclidean distance squared using norm $\|.\|$: if $i, i' \in \mathbb{R}^{|J|}$, i.e. these observations have values on attributes $j \in \{1,2,\dots,|J|\}$, J is the attribute set, $|.|$ denotes cardinality, then

$$d^2(i,i') = \|i - i'\|^2 = \sum_j (i_j - i'_j)^2. \quad (7)$$

Suppose that clusters C_i and C_j were next to be merged. At this point all of the current pairwise cluster distances are known. The recursive formula gives the updated cluster distances following the pending merge of clusters C_i and C_j. Let

- d_{ij}, d_{ik}, and d_{jk} be the pairwise distances between clusters C_i, C_j, and C_k, respectively,
- $d_{(ij)k}$ be the distance between the new cluster $C_i \cup C_j$ and C_k.

An algorithm belongs to the Lance-Williams family if the updated cluster distance $d_{(ij)k}$ can be computed recursively by

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|, \quad (8)$$

[a] Corresponding author: info@innovationlabs.net

where α_i, α_j, β and γ are parameters, which may depend on cluster sizes, that together with the cluster distance function determine the clustering algorithm.

3 Study of SMBs

In our study we analyzed data regarding subsidies provided by Moscow Small Business Administration in 2012 and 2013 to 810 companies. The size of the companies varies from 0 to 83 people, with the amount of cash available from 0 to 6,257,870 rubles (We must note that about half of the companies did not report any cash on hand. This probably indicates an issue with the data, rather than real information about the companies). The amount of subsidy received ranges from 123,109.68 rubles to 5,000,000.00 rubles and companies promised to create from 0 to 50 jobs for the money received. The aim of the study was to determine applicability of cluster analysis to the demonstration of the selection bias. The available data was Company ID, Company Size, Anticipated head count growth, Cash on hand, Subsidy size, Industry code, Final reporting. We applied Wards clustering algorithms to the data available and obtained 20 clusters varying in size from 1 to 187 companies (Table 1).

Table 1. Wards clustering of subsidy recipients. Cluster Sizes.

Cluster #	Cluster Size	Cluster #	Cluster Size
1	1	11	24
2	4	12	27
3	5	13	27
4	7	14	32
5	12	15	71
6	14	16	73
7	15	17	75
8	16	18	82
9	17	19	104
10	17	20	187

As we can see from the Table 1, the data forms 6 relatively large clusters (50+ entries), 10 medium size clusters (10-50 entries) and 4 small clusters (1-9 entries). For the propose of this study the latest 5 clusters form a group of a special interest as they naturally represent companies which are somehow different from the rest of the treated companies and can either demonstrate selection irregularities or some other special circumstances.

If we consider size of the treated companies and the proposed number of jobs created (Figure 1 and 2), we can see that most of the subsidy recipients have less than 20 people with two exemptions (cluster 2 and 4), and planning to create less than 20 jobs with one exemption (cluster 1). It seems to be clear that the only member of cluster 1 constitutes an irregularity, which could be attributed to the selection bias.

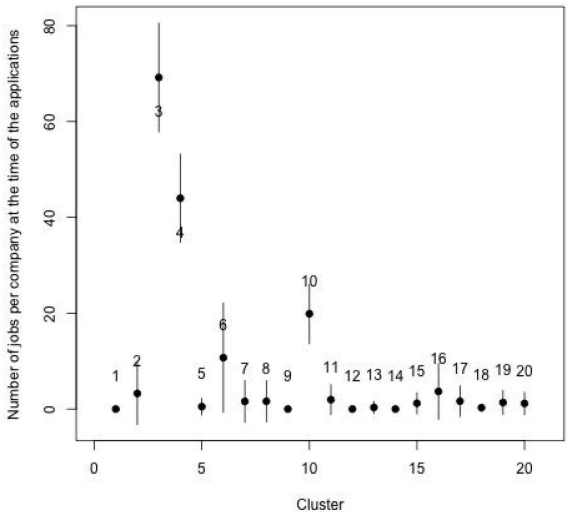

Figure 1. Number of jobs per company at the time of the applications.

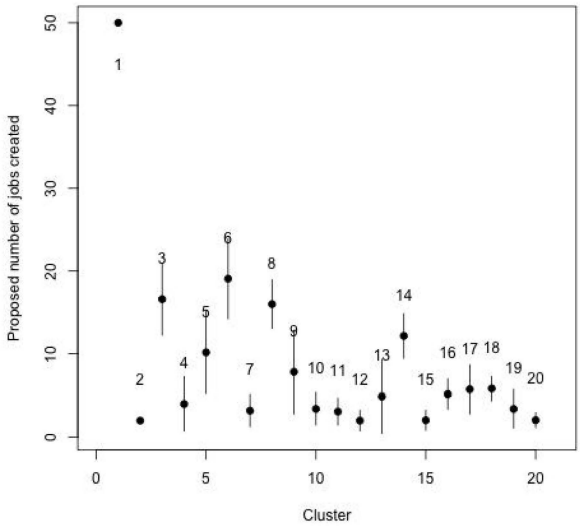

Figure 2. Proposed number of jobs created.

This conclusion becomes even clearer taken into account the amount of cash on hand at the time of application (Figure 3) and the amount of the subsidy received (Figure 4).

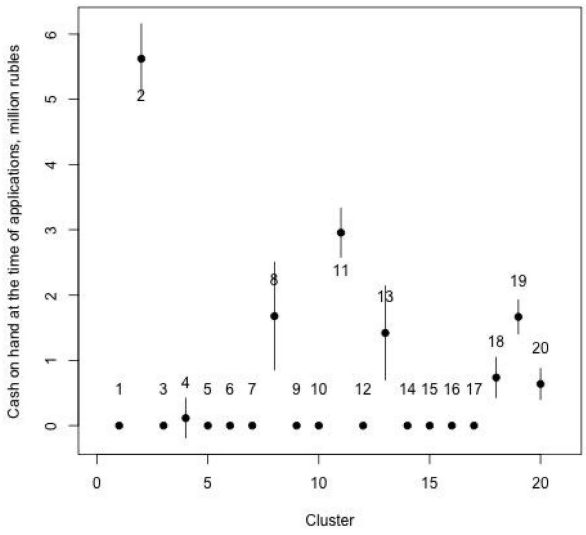

Figure 3. Cash on hand at the time of the applications.

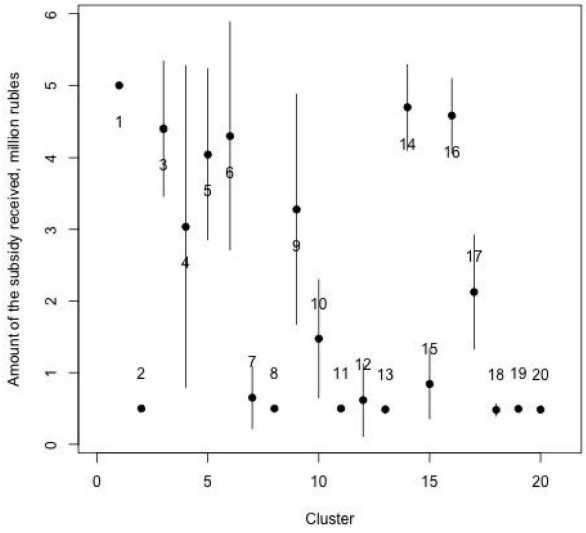

Figure 4. Subsidy received.

4 Conclusion

This paper demonstrates that even simple cluster analysis such as Ward's Hierarchical Agglomerative Clustering Method can be used to significantly reduce irregularities in selection of treatment recipients. Furthermore, we demonstrated that companies could be structured into a small number of clusters with similar properties, which later can be used to study treatment effect as suggested by Balk [4]. Unfortunately, the data available does not provide sufficient information to study treatment effect beyond proposed job creation, as data on actual job creation as well as IP creation was not part of recipients reporting at the time of this article's writing. Additionally, we have to note that data available for us is incomplete, as it did not contain industry information, and thus could not be directly used to find corresponding group of untreated companies.

References

1. D. Czarnitzki et al, ZEW Discussion Paper, *Evaluating the Impact of R&D Tax Credits on Innovation: A Microeconometric Study on Canadian Firms*, **04**, 77 (2004)
2. J. Hahn, *Identification and Estimation of Treatment Effects with Regression-Discontinuity Design*, Econometrica (2001)
3. M. Lechner, Journal of Business & Economic Statistics, *Earnings and Employment Effects of Continuous Off- the-Job Training in East Germany After Unification* (1999)
4. I. Balk, Economical Sciences, *Mathematical modeling of effectiveness of the government support of innovation using cluster modeling*, 5 (2015)
5. J.H. Ward jr., Journal of the American Statistical Association, *Hierarchical Grouping to Optimize an Objective Function*, 58 (1963)