Decision Memo for Rural Seed Sector Development

Through Participatory Varietal Selection

CS112: Knowledge: Information Based Decisions

Phuong Ha Tran Nguyen and Sona Vardanyan

Minerva Schools at KGI

Fall, 2020

To: Queensland Alliance for Statistical Methods in Agriculture

From: Haso Advising Agency

Date: 12/17/2020

Executive Summary

According to our analysis, we state that results gained by Wilkus, Francesconi, Jager (2017) can be more comprehensive through the use of better statistical tools. Our suggestions are:

- To use Genetic Matching as an additional method to verify their findings for *Unmatched Sample* and *Nearest Neighbor Matching.*
- To extend their existing study by performing a simulation and sensitivity analysis to gain more accurate results for both Genetic Matching and Nearest Neighbor Matching.

Problem and Objective

The first step of developing a client-oriented paradigm in agricultural development is to incorporate individual producers in development activities. However, most of the impact assessments remained focused on the ultimate outcomes, e.g. food security, poverty alleviation. Additionally, the empirical evidence showing the contribution of participatory activities in development agendas was lacking. The present study investigated the impact of participatory varietal section (PVS) trials on adoption and market participation.

While the ideal study is a Randomized Controlled Trial, in practice this was not achieved as the participants are part of existing programs, and tend to cluster around the same

block (Wilkus, et. al, 2017). Thus, control groups were chosen randomly but within similar geographical locations. This introduces many confounding variables, 10 of which are collected in the accompanied survey to assess the impact of PVS, namely: average age, the proportion of household members < 6, the proportion of household members < 60, household size, the proportion of household members that eat but do not contribute labor or financial support, proportion of household members that eat and provide labor and financial support, too weak to cultivate their fields, had to cultivate their fields alone, had the option of receiving extra help from paid labor, and had the option of receiving extra help from voluntary family members. Nearest-neighbor matching was used to ensure balance across all covariates.[1]

This decision brief aims to 1. Replicate the value achieved by the paper; 2. Introduce genetic matching as a superior method to achieve a better balance between covariates; 3. Use simulation to get a better estimate and confidence interval of treatment effect and 4. Conduct sensitivity analysis for the two findings after matching to find its robustness against unmeasured covariates.

Data and Methods

We start by loading and cleaning the data, as well as changing the treatment variable (affiliation) to a binary variable. We then replicate the columns "Unmatched Sample" and "Nearest Neighbor Matching" (NN) using a propensity score in table 4. The impact of the program on households was computed as the average treatment effect on the treated (ATT). We

[1] #variables: We have identified the independent (treatment) and dependent variables, as well as the covariates in the data in a systematic manner. We have used these variables respectively in the calculation and analysis of our methodology. We have also cross-checked with the paper to ensure that our understanding is correct.

confirm that prior to matching, the lowest p-value is < 0.001 for the variables paid_people and too_alone, but after matching, the lowest p-value is 0.29, which means that most variables are appropriately balanced because there is no significant difference in the means of the treatment and control groups. Furthermore, we find that the average distance for the propensity score between the control and treated groups reduces from 0.2158 to 0.0893, or a 58.64% improvement.[2]

We use Genetic matching to improve the results (see Appendix VII), which automatically finds the set of matches using genetic algorithms to optimize for the smallest discrepancy between the distribution of potential confounders in the treated and control group using the same basis of nearest neighbor pair matching (Noah 2020). Using genetic matching, we could achieve better overall distance and average difference, namely from 0.2158 to 0.0191, or a 91.16% improvement.

It is expected that with better matching, we would be able to get a better estimate of ATT. Moreover, the treatment effect can be more representative and clear as a 95% confidence interval instead of a single point estimate. However, since the treatment effect consists of 5 intended consequences and 4 unintended consequences, we only perform the treatment effect estimate for the first intended consequence i1, which is Receiving subsidized or free seeds. We further calculate both the expected value, which is when the simulations only include estimation uncertainty, and predicted values, when the simulations take into account both estimation uncertainty (due to sampling effect) and fundamental uncertainty (irreducible errors due to

[2] #studyreplicatoin: We use an existing study by Wilkus (2017), specifically Table 4 from that paper. Firstly use their method to replicate the results and improve it by using Genmatch and Sensitivity Analysis.

different individuals' context). The former is more useful in causal inference, and the later in prediction models. Since we are understanding the overall average treatment effect of PVS, expected values would be more useful. We find using NN that the mean for ATT is 0.509, and its 95% confidence interval is [0.4419819, 0.5688817] (Appendix V). Using genetic matching, the mean for ATT is 0.596, and its 95% confidence interval is [0.518, 0.671] (Appendix VIII). Both suggest that the treatment has a significant effect on an individual's ability to buy seeds at subsidized or free rates. This is an illustration of the method which can be extended further for other intended and unintended outcomes in the paper. The expected value and its confidence intervals should allow policymakers to understand the effect of PVS intuitively and investigate more resources accordingly.[3]

Finally, we ran Rosenbaum's sensitivity analysis for the model for both genetic matching and NN. This is necessary to account for the robustness of the current model against unmeasured covariates, which is highly likely in this case due to the quasi-experiment design of the study. After matching using NN, we find that its robustness gamma is 6.6, which means that one subject needs to be 6.6 times more likely than another to receive treatment due to the differences in unobserved covariates before the conclusion of the study would change. With any hidden covariates that make the odds smaller, the effect would still be statistically significant at the 95% confidence level. Meanwhile, using genetic matching, the robustness gamma is 4.3, which is smaller than matching using NN, and suggests that the model, although better matched, is less

[3] #correlation: Through matching, we can remove the confounding variables that would differentiate a correlation versus causation. Furthermore, using a confidence interval for the treatment effect, we can be confident that there is a significant difference between treatment and control group, which can only be attributed to the treatment variable.

robust to hidden bias. Nonetheless, both values suggest that the model must be relatively robust to bias, because it is unlikely that any combination of hidden variables that were not recorded by the survey could have such a substantial impact on the odds of receiving treatment.

Conclusion and Suggestions

Because there is a trade-off between accuracy of matched variables and robustness against hidden bias, we are unable to recommend genetic matching as a superior method to nearest neighbor matching. Nonetheless, we find it useful to simulate the results and examine the hidden bias for both methods of matching and we recommend its usage for a better estimation of the causal effect of PVS on receiving free and subsidized seeds. We further recommend for the methods to be replicated for the remaining treatment effects, and/or for a multivariate outcome variable to be established such that we can better understand the overall effect of PVS on rural seed sector development.

Appendix

A. R Coding

Also available here: https://rpubs.com/sherlockieee/finalproject

```
#get required library
library(readxl)
library(MatchIt)
library(Matching)
```

```
library(rbounds)
library(MASS)
library(dplyr)
```

```
library(janitor)
```

```
library(cobalt)
```

```
#devtools::install_github("Ngendahimana/sensitivityR5", force=TRUE)
library(SensitivityR5)

set.seed(123)
```

```
#load dataframe
df <- read_excel("02.SeedSectorData.xlsx") %>% clean_names() %>% remove_empty()
```

I.

```
#change treatment variable to boolean
df$treatment = ifelse(df$affiliation == "Af", 1, 0)
head(df)
```

```
## # A tibble: 6 x 27
##    hhid seed_collection fgroup affiliation    ky    ka    ak av_age under6
##   <dbl>           <dbl> <chr>  <chr>       <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1     1               0 Ky     Af              1     0     0   56.7  0
## 2     2               1 Ky     Af              1     0     0   19.6  0
## 3     3               1 Ky     Af              1     0     0   31.6  0
## 4     4               1 Ky     Af              1     0     0   22.2  0.333
## 5     5               0 Ky     Af              1     0     0   18.9  0
## 6     6               0 Ky     Af              1     0     0   18.3  0.333
## # … with 18 more variables: over60 <dbl>, hh_size <dbl>, prop_eat <dbl>,
## #   prop_all <dbl>, too_weak <dbl>, too_alone <dbl>, paid_people <dbl>,
## #   volunteers <dbl>, i1 <dbl>, i2 <dbl>, i3 <dbl>, i4 <dbl>, i5 <dbl>,
## #   u1 <dbl>, u2 <dbl>, u3 <dbl>, u4 <dbl>, treatment <dbl>
```

II.

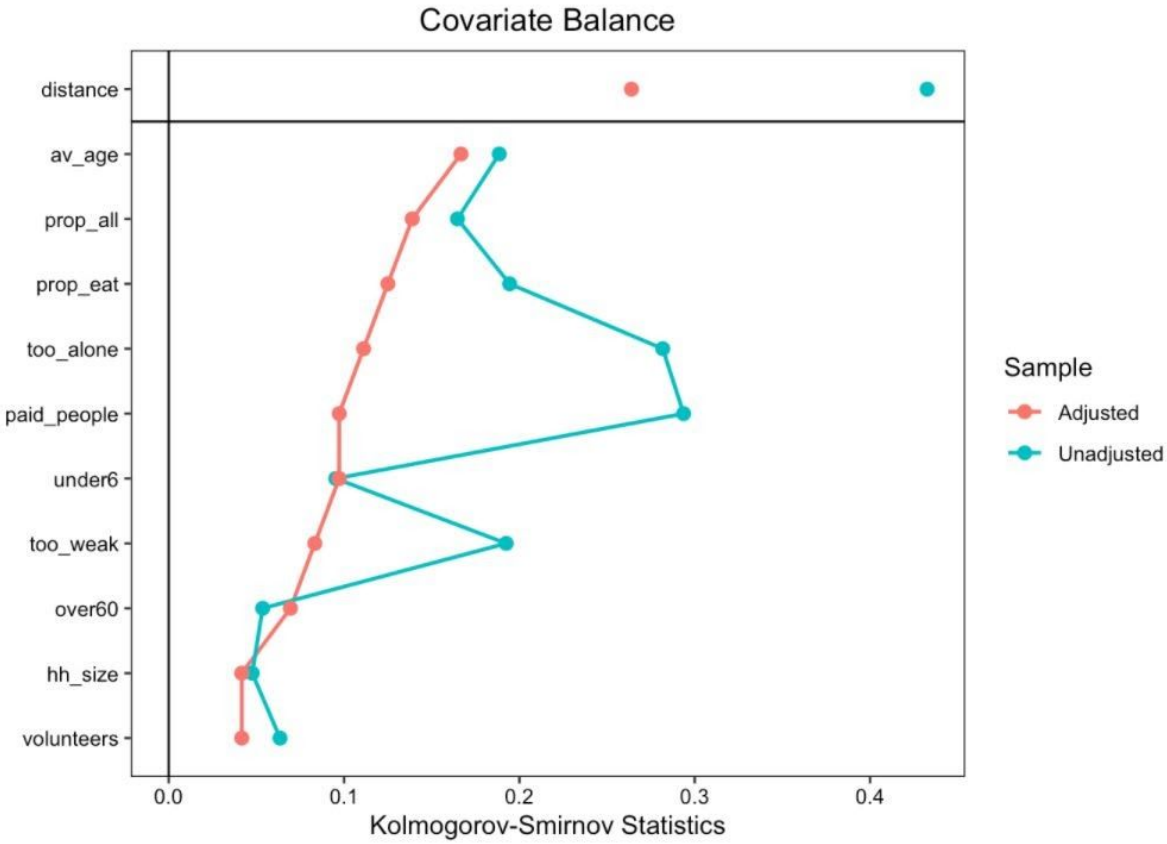
```
#using nearest neighbor matching
nn.mout <- matchit(treatment ~ av_age + under6 + over60 + hh_size + prop_eat + prop_all + too_weak + too_alone + paid_people
+ volunteers, data = df, estimand = "ATT", method = "nearest")

#print matching improvement
print(bal.tab(nn.mout))
```

```
## Call
##  matchit(formula = treatment ~ av_age + under6 + over60 + hh_size +
##     prop_eat + prop_all + too_weak + too_alone + paid_people +
##     volunteers, data = df, method = "nearest", estimand = "ATT")
##
## Balance Measures
##                Type Diff.Adj
## distance    Distance   0.4283
## av_age       Contin.  -0.1286
## under6       Contin.   0.0057
## over60       Contin.  -0.1853
## hh_size      Contin.   0.0064
## prop_eat     Contin.   0.0241
## prop_all     Contin.  -0.0025
## too_weak      Binary   0.0833
## too_alone     Binary   0.1111
## paid_people   Binary  -0.0972
## volunteers    Binary  -0.0417
##
## Sample sizes
##           Control Treated
## All           126      72
## Matched        72      72
## Unmatched      54       0
```

III.

```
#plot balance between covariates before and after matching
love.plot(nn.mout, abs=TRUE, var.order = "adjusted", line = TRUE, stats= c("ks.statistics", "abs difference"))
```



IV.

```
#calculate treatment effect using simulation
z.out <- zelig(i1 ~ treatment + av_age + under6 + over60 + hh_size + prop_eat + prop_all + too_weak + too_alone + paid_peopl
e + volunteers, data = match.data(nn.mout), model = "ls")
```

```
## Warning: `tbl_df()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## Warning: `group_by_()` is deprecated as of dplyr 0.7.0.
## Please use `group_by()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## How to cite this model in Zelig:
##   R Core Team. 2007.
##   ls: Least Squares Regression for Continuous Dependent Variables
##   in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,
##   "Zelig: Everyone's Statistical Software," https://zeligproject.org/
```

V.

```
x.out <- setx(z.out, data = match.data(nn.out, "treat"), cond = TRUE)
s.out <- sim(z.out, x = x.out)

summary(s.out)
```

```
##
##  sim x :
##  -----
## ev
##        mean         sd       50%      2.5%    97.5%
## 1 0.5057401 0.03231898 0.5043022 0.4470116 0.570547
## pv
##           mean        sd       50%       2.5%    97.5%
## [1,] 0.4797118 0.3974771 0.4780563 -0.2896219 1.256887
```

VI.

```
#using genetic matching
gen.mout <- matchit(treatment ~ av_age + under6 + over60 + hh_size + prop_eat + prop_all + too_weak + too_alone + paid_peopl
e + volunteers, data = df, estimand = "ATT", method = "genetic", print.level = 0, pop.size = 1000, unif.seed=112,  int.seed=
112)
```

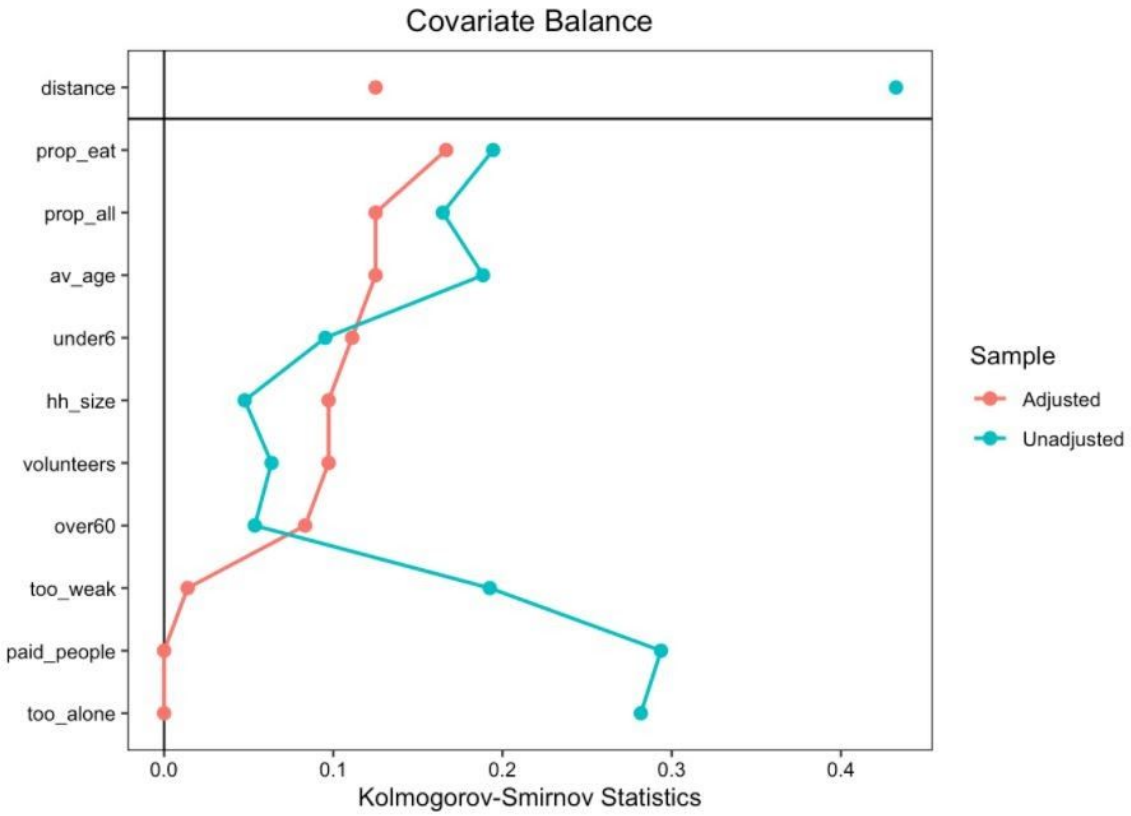
```
## Loading required namespace: rgenoud
```

```
#print matching improvement
print(bal.tab(gen.mout))
```

```
## Call
##  matchit(formula = treatment ~ av_age + under6 + over60 + hh_size +
##     prop_eat + prop_all + too_weak + too_alone + paid_people +
##     volunteers, data = df, method = "genetic", estimand = "ATT",
##     print.level = 0, pop.size = 1000, unif.seed = 112, int.seed = 112)
##
## Balance Measures
##                 Type Diff.Adj
## distance    Distance   0.0915
## av_age       Contin.   0.0796
## under6       Contin.   0.0354
## over60       Contin.   0.1441
## hh_size      Contin.   0.0191
## prop_eat     Contin.   0.0830
## prop_all     Contin.   0.1455
## too_weak      Binary   0.0139
## too_alone     Binary   0.0000
## paid_people   Binary   0.0000
## volunteers    Binary  -0.0972
##
## Sample sizes
##                       Control Treated
## All                    126.        72
## Matched (ESS)           25.66      72
## Matched (Unweighted)    40.        72
## Unmatched               86.         0
```

VII.

```
#plot balance between covariates before and after matching
love.plot(gen.mout, abs=TRUE, var.order = "adjusted", line = TRUE, stats = "ks.statistics")
```



VIII.

```
#calculate treatment effect using simulation
gen.z.out <- zelig(i1 ~ treatment + av_age + under6 + over60 + hh_size + prop_eat + prop_all + too_weak + too_alone + paid_p
eople + volunteers, data = match.data(gen.mout), model = "ls")
```

```
## How to cite this model in Zelig:
##   R Core Team. 2007.
##   ls: Least Squares Regression for Continuous Dependent Variables
##   in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,
##   "Zelig: Everyone's Statistical Software," https://zeligproject.org/
```

```
gen.x.out <- setx(gen.z.out, data = match.data(gen.mout, "treat"), cond = TRUE)
gen.s.out <- sim(gen.z.out, x = gen.x.out)

summary(gen.s.out)
```

```
##
##  sim x :
##  -----
## ev
##       mean        sd       50%      2.5%     97.5%
## 1 0.596518 0.0392707 0.5966841 0.5195656 0.6720248
## pv
##           mean        sd       50%       2.5%    97.5%
## [1,] 0.5863244 0.3959473 0.5731946 -0.1884281 1.339729
```

IX.

```
#sensitivity analysis for genetic matching
pens2(x = gen.mout, y="i1",Gamma = 5, GammaInc = 0.1, est = 0.5987271)
```

```
##
##  Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-Value
##
## Unconfounded estimate ....  0
##
##  Gamma Lower bound Upper bound
##    1.0           0      0.0000
##    1.1           0      0.0000
##    1.2           0      0.0000
##    1.3           0      0.0000
##    1.4           0      0.0000
##    1.5           0      0.0000
##    1.6           0      0.0000
##    1.7           0      0.0001
##    1.8           0      0.0001
##    1.9           0      0.0002
##    2.0           0      0.0003
##    2.1           0      0.0005
##    2.2           0      0.0008
##    2.3           0      0.0011
##    2.4           0      0.0015
##    2.5           0      0.0021
##    2.6           0      0.0028
##    2.7           0      0.0037
##    2.8           0      0.0047
##    2.9           0      0.0060
##    3.0           0      0.0074
##    3.1           0      0.0091
##    3.2           0      0.0109
##    3.3           0      0.0131
##    3.4           0      0.0154
##    3.5           0      0.0181
##    3.6           0      0.0210
##    3.7           0      0.0241
##    3.8           0      0.0276
##    3.9           0      0.0313
##    4.0           0      0.0352
##    4.1           0      0.0395
##    4.2           0      0.0439
##    4.3           0      0.0487
##    4.4           0      0.0537
##    4.5           0      0.0590
##    4.6           0      0.0645
##    4.7           0      0.0702
##    4.8           0      0.0762
##    4.9           0      0.0824
##    5.0           0      0.0888
##
##  Note: Gamma is Odds of Differential Assignment To
##  Treatment Due to Unobserved Factors
##
```

X.

```
#sensitivity analysis for nearest neighbor matching
pens2(x = nn.mout, y="i1",Gamma = 7, GammaInc = 0.1, est = 0.5057401)
```

```
##
##  Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-Value
##
## Unconfounded estimate ....  0
##
##  Gamma Lower bound Upper bound
##    1.0           0      0.0000
##    1.1           0      0.0000
##    1.2           0      0.0000
##    1.3           0      0.0000
##    1.4           0      0.0000
##    1.5           0      0.0000
##    1.6           0      0.0000
##    1.7           0      0.0000
##    1.8           0      0.0000
##    1.9           0      0.0000
##    2.0           0      0.0000
##    2.1           0      0.0001
##    2.2           0      0.0001
##    2.3           0      0.0001
##    2.4           0      0.0002
##    2.5           0      0.0003
##    2.6           0      0.0004
##    2.7           0      0.0005
##    2.8           0      0.0006
##    2.9           0      0.0008
##    3.0           0      0.0010
##    3.1           0      0.0013
##    3.2           0      0.0016
##    3.3           0      0.0019
##    3.4           0      0.0023
##    3.5           0      0.0028
##    3.6           0      0.0033
##    3.7           0      0.0039
##    3.8           0      0.0045
##    3.9           0      0.0052
##    4.0           0      0.0059
##    4.1           0      0.0068
##    4.2           0      0.0076
##    4.3           0      0.0086
##    4.4           0      0.0096
##    4.5           0      0.0107
##    4.6           0      0.0119
##    4.7           0      0.0131
##    4.8           0      0.0144
##    4.9           0      0.0158
##    5.0           0      0.0172
##    5.1           0      0.0188
##    5.2           0      0.0204
##    5.3           0      0.0220
##    5.4           0      0.0237
##    5.5           0      0.0255
##    5.6           0      0.0274
##    5.7           0      0.0293
##    5.8           0      0.0313
##    5.9           0      0.0334
##    6.0           0      0.0355
##    6.1           0      0.0377
##    6.2           0      0.0399
##    6.3           0      0.0422
##    6.4           0      0.0446
##    6.5           0      0.0470
```

```
##    6.6          0      0.0495
##    6.7          0      0.0520
##    6.8          0      0.0545
##    6.9          0      0.0572
##    7.0          0      0.0598
##
##  Note: Gamma is Odds of Differential Assignment To
##  Treatment Due to Unobserved Factors
##
```

Reference

Greifer, N. (2020, December 15). Matching Methods. Retrieved December 18, 2020, from https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html

University College London. (n.d.). Introduction to Quantitative Methods. Retrieved December 18, 2020, from https://www.ucl.ac.uk/~uctqiax/PUBLG100/2016/faq/zelig.html

Wilkus, E.L., Francesconi, G.N. and Jäger, M. (2017), "Rural seed sector development through participatory varietal selection: Synergies and trade-offs in seed provision services and market participation among household bean producers in Western Uganda", Journal of Agribusiness in Developing and Emerging Economies, Vol. 7 No. 2, pp. 174-196. https://doi.org/10.1108/JADEE-01-2016-0002