Impact of Financial Access on School Attendance

Group No. 5

February 11, 2019

Members:

- Hector Hernandez Heimpel (NetID: hh744) - email: hh744@georgetown.edu
- Tingjie Meng (NetID: tm1305) - email: tm1305@georgetown.edu
- Liumin Chen (NetID: lc1077) - email: lc1077@georgetown.edu

Github Repository: https://github.com/hheimpel/Group-No.-5

Problem Statement and Background

There are multiple data sets which will comprise the source of our analysis. We will be using the Mexican Family Life Survey (ENNVIH) as the main source. The ENNVIH consists of multiple books each containing a set of data sets in `.dta` form; the first group of data sets is contained in the Control Book for 2002. The data sets contained within this book contain household level characteristics and they are divided into two categories. The first category contains characteristics observed in the household by the interviewer (i.e. proximity to water source, type of floor, etc.) and the second category contains answers to individual level survey questions (i.e. marital status, maximum education grade, etc.). The second group of data sets are contained in Book V for 2002 in the Mexican Family Life Survey. These group of datasets contain information about characteristics of children under 15 years of age, they include but are not limited to, child employment, child health characteristics and more.

Methods

Tools

Constructing Data for 2002 & 2005

Manipulating the data and preparing it for analysis required a great deal of work; we describe the process, the main tools used to clean the data and all of the manipulations done in order to obtain the final working data sets.

Household Level Data for 2002 & 2005

We start by loading the control book for 2002 (M X F L S - Book C). This book contains several control characteristics at an individual level and at a household level. Within the book there are four pertinent data sets to be used:

1. `c_portad.dta`: contains the location of the households by State, Municipality and Locality. It further contains a variable indicating number of inhabitants in Locality. This is a Household level data set.
2. `c_cv.dta`: contains physical characteristics of the household (i.e. whether it has a cooking room, a telephone, etc.). This is a Household level data set.
3. `c_cvo.dta`: this data set extends the previous one with more physical characteristics of the households (i.e. the construccion materials, access to electricity, etc.). This is a household level data set.
4. `c_ls.dta`: it contains control characteristics such as income, level of education, gender, among others. This is an individual level data set.

Since the data sets were in `dta` form, the `haven` package was required to load them into the environment and the package `here` was used was well for reproducibility purposes. Furthermore, use of the `tidyverse` package was used extensively throughout the process.

Once the data sets were loaded we proceeded to eliminate several non-informative variables. Some of them were redundant variables such as more identification characteristics for the household members, other variables seemed to ask the same thing but with a slightly different wording. For more details please check the Codebook for book C annexed in the project files.

The next step consisted of recoding the values of certain variables to reflect the different categories as the codebook instructed. The codebook provided numeric values to categorical variables which could be slightly confusing, we therefore recoded the values to reflect in a string what they actually meant. Additionally, we recoded dummy variables to be equal to one if the condition was true and zero otherwise (they were previously coded as threes as opposed to zeros); lastly, we created a resumed categorical variable for the size of the locality, if population was bigger thatn 100 thousand we categorized as `urban` and `rural` otherwise.

In many of the books many missing values for certain variables were observed. We created a data frame per data set to show us how many missing values each variable had and kept only the variables which didn't have an excessive amount of missing values. The first data set we noticed a missing values problem was `c_cv`. The second data set where we applied the same logic was `c_cvo`; we removed the variables which have over forty percent of missing values in these data sets.

Once we kept the variables with no issues in all the data sets previously mentioned we merged the data sets into a master data set. It's important to mention that the merging was done by the variable `folio` which is the individual household id; by doing so, there were several household level characteristics which were attributed to an individual level. Observations of the households contruction materials for example will be the same for each individual member of the household. Finally, some recoding was done to as before to reflect the true categoreis of a variable with strings and dummies were appropriately transformed.

It is worth mentioning the transformation of three individual level variables to household level variables. The first variable measured the highest education grade the household member had attended; since the focus will be children between 6 and 15 years of age we transformed the variable to show the average **adult** household member education. It seems likely that children whose accompanying adult household member have a higher education level will be more likely to have more education.

The second variable we transformed was income. Since income only reflected the individual level of income and since many children do not have income themselves we transformed the variable to show the average income per household. Intuitively, if the aggregated or average income per household is bigger then perhaps children have more opportunities to attend school. It would have seemed naive to impute missing values of income for children with things other than a household level transformation of the individual level variable income.

The third variable we transformed was relationship status. Since it is safe to assume that most children under 15 are single we decided to create a household level variable which reflected the relationship status of the household head. The intuition behind this is that perhaps there is some correlation between having married figures as parents or household heads.

Once we created all the household level transformations to the mentioned variables, we filtered the data to only include children older than 5 and younger than 15.

Child Individual Data

We then proceeded to load individual level data for children by loading the book V for 2002 (M X F L S - Book V). This book contains several characteristics at an individual level for children under 15 years of age. Within the book there are six pertinent data sets to be used:

1. `v_edna.dta`: contains information regarding children's education.
2. `v_emn.dta`: contains information regarding child labor.
3. `v_cen.dta`: contains information regarding outpatient utilization for children.
4. `v_atn.dta`: contains information regarding time allocation for children.
5. `v_esn.dta`: contains information regarding overall children's health.
6. `v_hsn.dta`: contains information regarding inpatient utilization for children.

These data sets contained many variables with excessive amounts of missing values. Once again we went through all data sets to identify variables which had excesssive amounts of missing values.

Once we identified the variables with many missing values we then proceeded to discard redundant variables and to merge all of the data sets into a master data set. Categorical variables were also transformed to show the actual string category they represent as opposed to numeric value. The last step consisted of recoding dummy variables to have a value of one if condition was true and zero otherwise.

Merging Book V and Book C master data sets

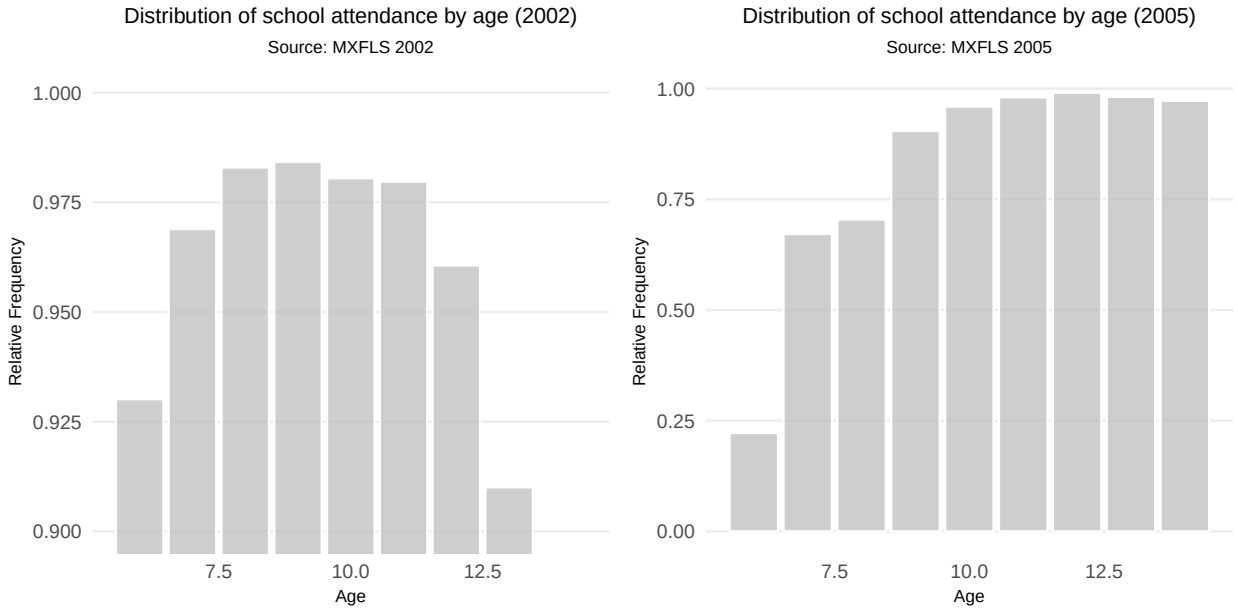
To have the working master data set we merged both of the master data sets from book C and book V. Lastly, we imputed household level missing variables (missing observations at a household level will be substituted for whichever value or characteristic other members of the household have).

We repeat the same process as for the 2002 data set to create the 2005 data set. All the data sets and variables are named the same and the same patterns of missing values were displayed in 2005 as in 2002; this made it a lot easier to replicate the code with minor changes.

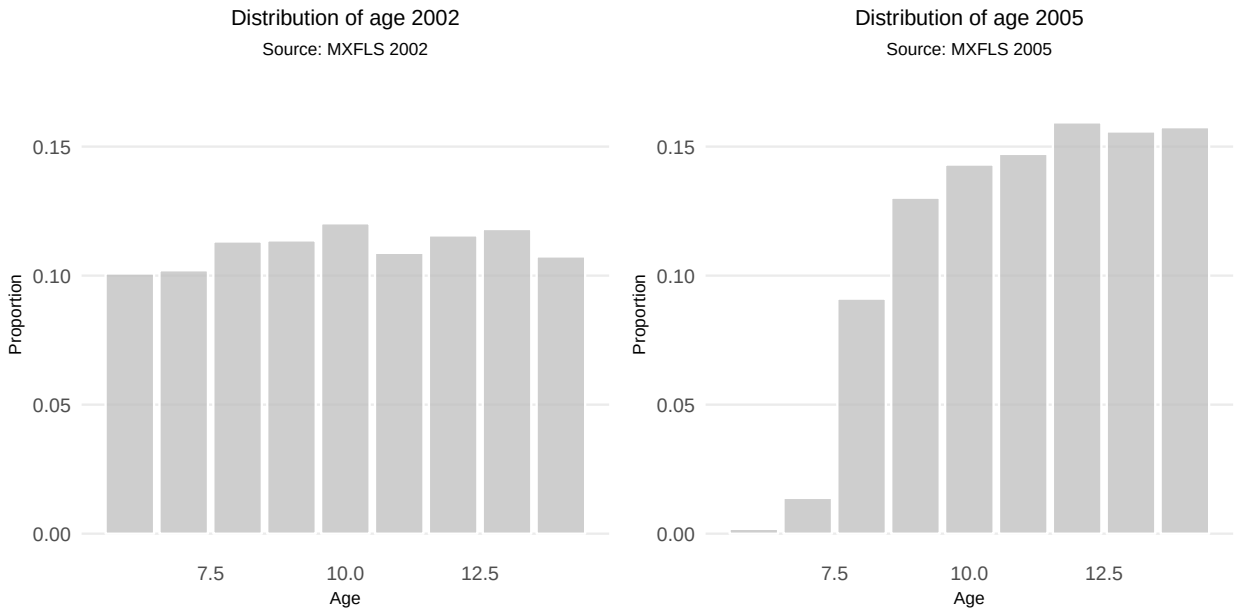
Exploratory Analysis

School Attendance Distribution

In our exploratory analysis, we started by simply looking at School Attendance distribution by age in 2002 and 2005:



There seemed to be some fundamental change from 2002 to 2005 in school attendance, especially for young children under the age of seven. We explored the missing values in both data sets to look for systematic patterns of missing data and found that over 700 missing values for the variable `age` were found in the 2002; in contrast, over 3,500 (almost 40 percent of the observations) missing values were found in the 2005 data set. We cannot ignore the systematic number of age missing values and perhaps this is what is causing the dramatic shift in school attendance from 2002 to 2005. Since this is panel data, a plausible hypothesis in the data recollection process is that new members of the household could have been added but for the household control questions in 2005 they only asked the information for the people present in 2002. The following graphs compare the age distribution across both data sets:

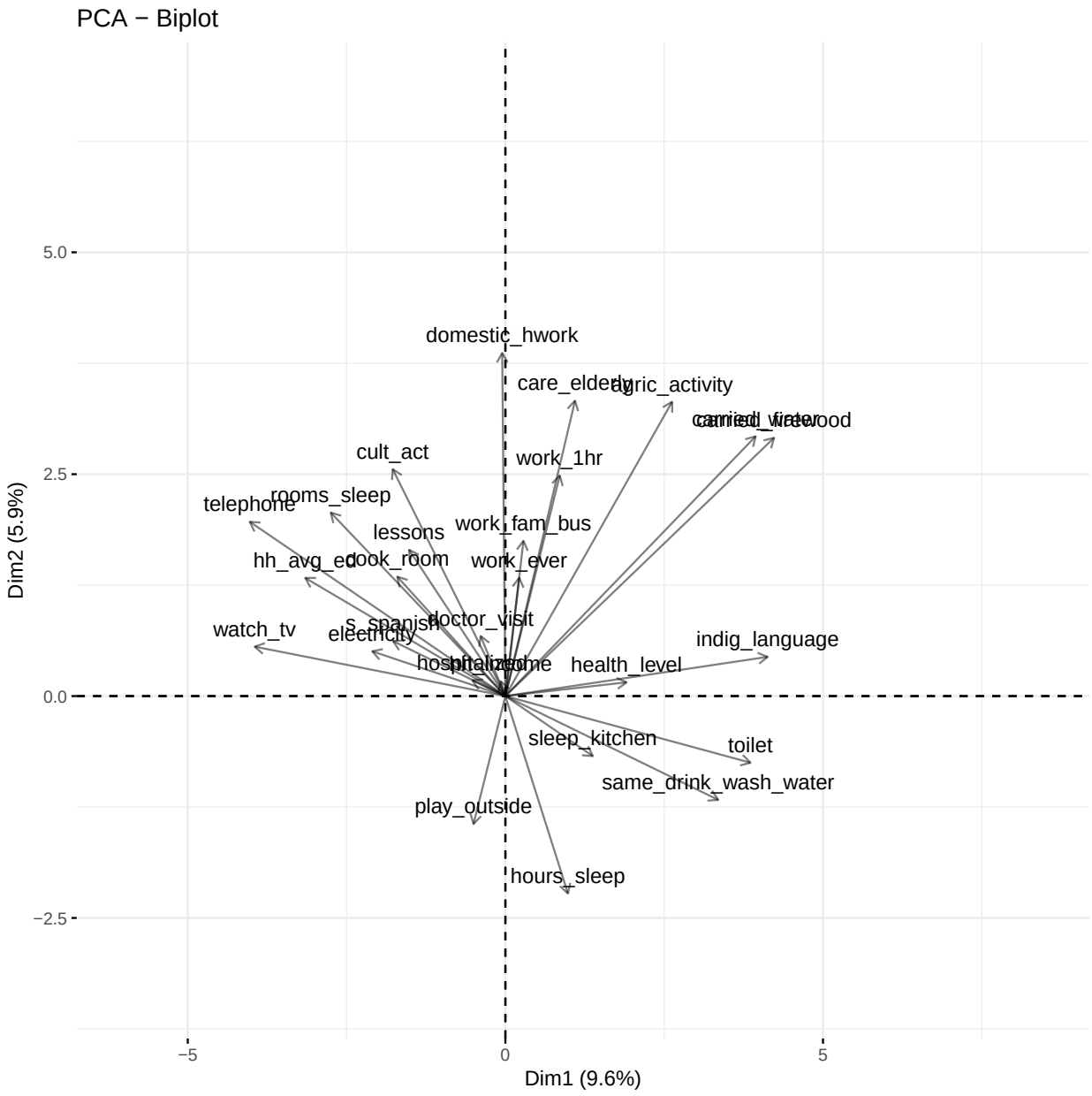


The graph seem to corroborate the hypothesis that in the data collection process they focused on interviewing people who were present in the 2002 round of surveys. This could potentially have some influence in the results so it is important to keep this in mind if the models are to be applied somewhere else.

Principal Component Analysis

Since we have many covariates in our data set, we looked at the correlation matrix to detect any potential problems with the variables (To look at the correlation plot refer to Appendix - Part 1). We looked at numeric variables and dummies first and there were two variables which seem to have issues and it's intuitive to ascertain why. The correlation between the variable `work_1hr` indicating whether the child worked at least one hour the past week, `work_ever` which indicates whether the child has ever worked and `work_fam_bus` is missing; this could be due to the fact that they seem to be measuring the same thing. There seem to be fairly small numbers for the correlation of variables; however, this is not the entirety of the story, underlying structures could be hidden.

To try to find underlying hidden structures in the data, we resorted to principal component analysis to see if the number of variables in the data could be reduced; for this purpose we used the `factoextra` and `FactoMineR` packages. The first step would be transforming categorical variables into dummy variables; however, given the number of variables this would result in, we focus on numerical and dummy variables at first. The second step is removing identification variables and the third step is scaling the remaining variables. The result is the following:



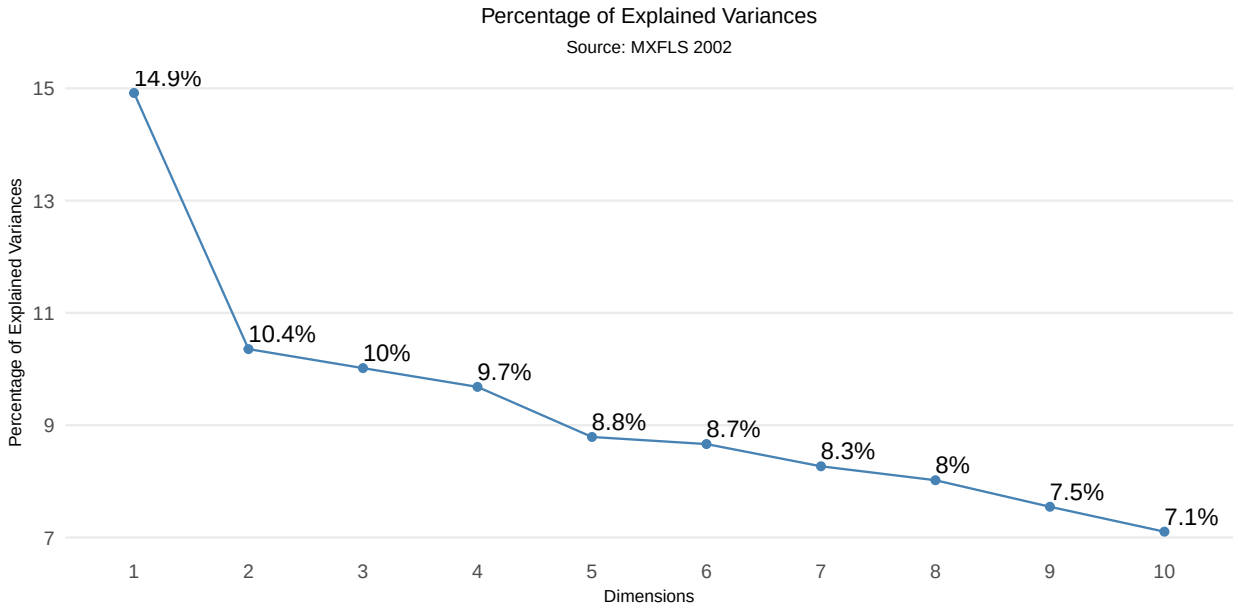
Some very interesting results can be observed from the previous graph:

1) The vertical dimension can almost be entirely classified by the variable `domestic_hwork` which refers to domestic work in the household.
2) The horizontal dimension includes several household characteristics. This dimension seems to suggest something along the lines of a welfare measure.
3) The variables `work_1hr`, `work_fam_bus`, `work_ever` and `care_elderly` are very close together and near the vertical dimension. This suggest that they are measuring the same work dimension.
4) The variables `electricity`, `watch_TV` and `telephone` are close in space. This makes intuitive sense because to have either a telephone or a TV you must have electricity.
5) The variables `play outside` and `hours_sleep` seep opposite to the vertical dimension

which seems to be something relating to work; this makes intuitive sense because likely children who do domestic work sleep less and play less.

6) The variables `sleep_kitchen`, `same_drink_wash_water` and `toilet` are also very close in space. It makes sense because household where someone sleeps in the kitchen, the same water is used for washing and drinking and there are no toilets could be measuring something along the lines of household quality.
7) `carried_firewood` and `carried_water` seem to be measuring the same thing, carrying a production input or something along these lines.
8) The number of rooms to sleep `rooms_sleep` and `cook_room` which are in close opposite direction of variables in point 6. are also very close together.

We proceeded to filter out these variables (keeping a single one for reference) to condense our number of variables. The results are the following:



The remaining dimensions have a very uniform percent of explained variance; the range for the maximum and minimum for the eigenvalues is (7.1%, 14.9%). This seems to suggest that we have reduced the dimensions as much as possible.

We repeated the same principal component analysis (PCA) for categorical variables (refer to Appendix - Part 2). The `caret` package was used to transform categorical variables into dummies to make the analysis possible. Given the many categorical variables and the many categories within those variables we first analyzed those categories which we intuitively thought would be most related. The first category was "physical characteristics of the household" (i.e. material of the walls, materials of the floor, access to water, etc.). From rounds one through five we applied the PCA for each categorical variable in the broader category of physical characteristics of the household and eliminated the variables which provided the least in terms of percentage of explained variance; for round six we compiled all of the relevant categories from the previous rounds and ran PCA again; finally, in round 7 we kept the six categories for physical characteristics of the household which provided the

most in terms of explained varianc The same process was repeated for categorical variables `marriage_status` (i.e. single, married, divorced, etc.) and `property_status` (i.e. rent, own, communal land, etc.). From over 90 covariates we managed to reduce the number to 30.

Supervised Learning

Once the covariates were significantly reduced we proceeded to build a a supervised learning model to ascertain which covariates had the most predictive power for school attendance. Given that the rate of school attendance is over 90%, our measure of choice for the goodness of fit of the model was Specificity (of the actual true negatives, how many were correctly classified). The first step was creating a training data set and a testing data set and rescaling continuous variables by applying logarythms.

The next step consisted of preparing the data set by normalizing continuous variables (categorical variables were previously converted to dummies). Furthermore, missing values were imputed (using the mode for dummy variables and the median for continuous variables). Next, we set the number of cross-validations to 5 and since this is a classification problem, the proper settings were established. The `recipes` package was necessary for these steps.

We then ran a CART model and tuned the complexity parameter until we found results that converged to a maximum level of specificity (for details check Appendix - Part 3).

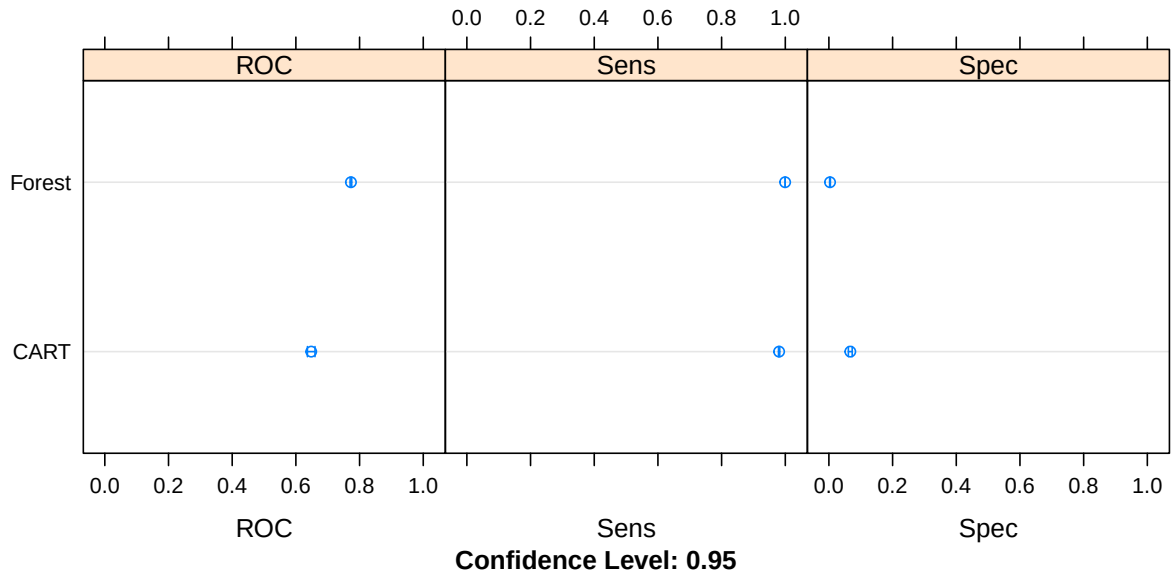
The second model we ran was a random forest, the packages `ranger` and `e1071` were needed for this purpose. The tuning parameter was `mtry` which states how many predictors the model can take (for details check Appendix - Part 3).

Banks Data

Once we reduced the number of covariates considerably, we proceeded to create the bank information data set. From an xls spreadsheet in the Comision Nacional Bancaria y de Valores (CNBV) we loaded the data set called `BM_Operativa_200212.xls` using the `readxl` package; this data set contains the information aboout the number of bank branches per municipality in Mexico (with an exhaustive list of every bank authorized for operations in Mexico by the CNBV). Extensive string manipulations with the `stringr` package had to be done in order to make the data sets comparable with those in the previous sections; this was due to the fact that the locations were encoded as strings as opposed to numbers and even with another data set called `Relacion_de_municipios_de_Mexico.xlsx` which contains the relationship between municipalities names and their official corresponding numbers, the merging process required extensive manipulation.

Results

We started by examing the supervised learning models to assess which one had the highest specificity level. Even though specificity was low, we can observe from the following graph that the CART model had the highest level. This was corroborated by testing the model in out of sample data Appendix - Part 3).



We chose the CART model to obtain information about the predictive accuracy for the variable `attend_school` of each covariate. We conjectured that the top 6 covariates were enough controls to include in our linear regression, these are:

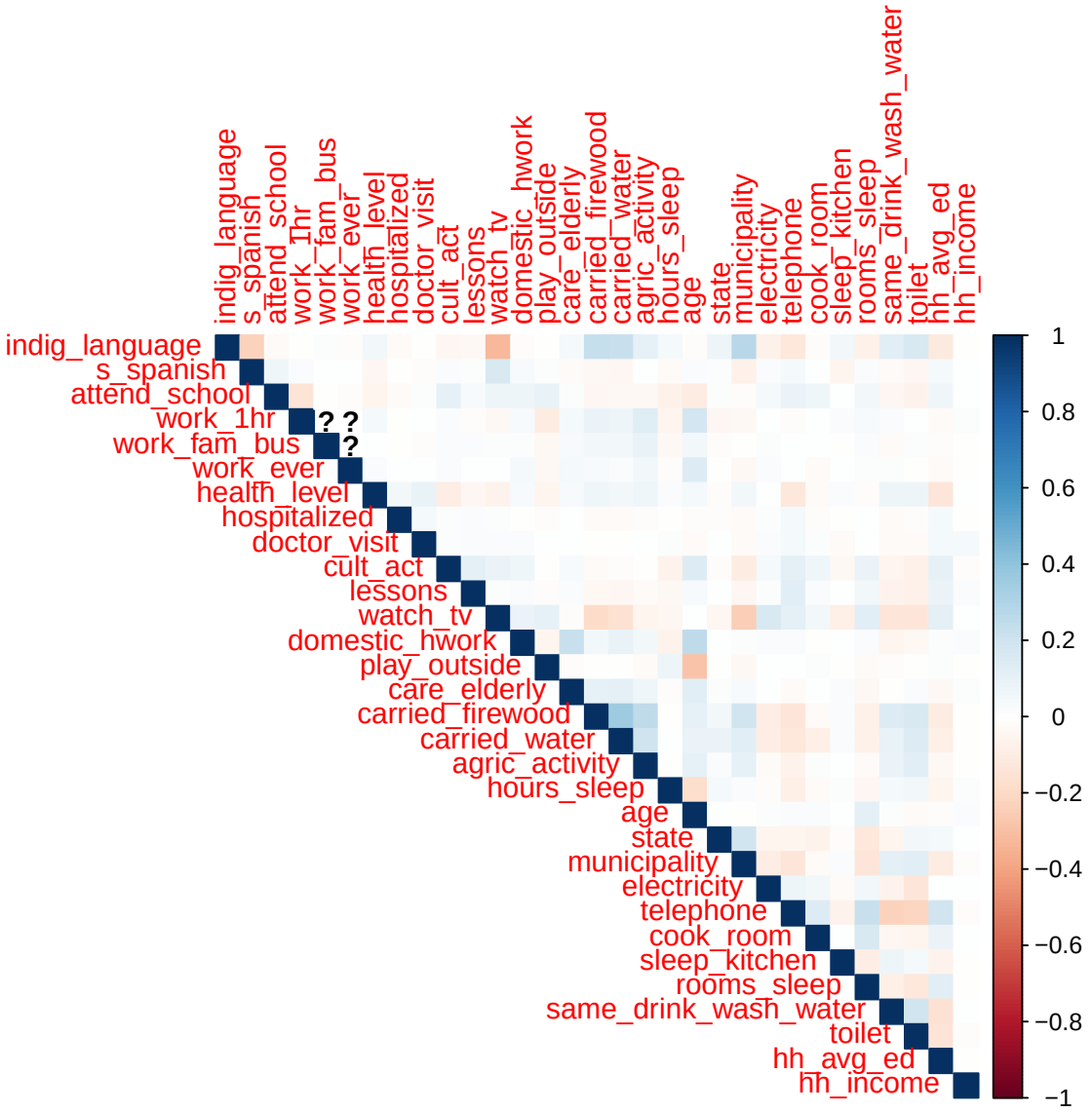
- `age`: age of the child at time of interview.
- `hours_sleep`: average sleeping hours for child.
- `hh_avg_educ`: average level of education of household members.
- `hh_income`: household combined income.
- `cult_act`: whether ther have performed recent cultural activites.
- `domestic_hwork`: whether the child helps in domestic household work.

References

- Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra
- Brandon M. Greenwell (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421–436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.
- Hadley Wickham and Evan Miller (2018). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.0.0. https://CRAN.R-project.org/package=haven
- Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
- Kirill Müller (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here
- Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01
- Max Kuhn and Hadley Wickham (2019). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.5. https://CRAN.R-project.org/package=recipes
- Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-82. https://CRAN.R-project.org/package=caret
- Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

Appendix

Part 1 - Correlation Matrix

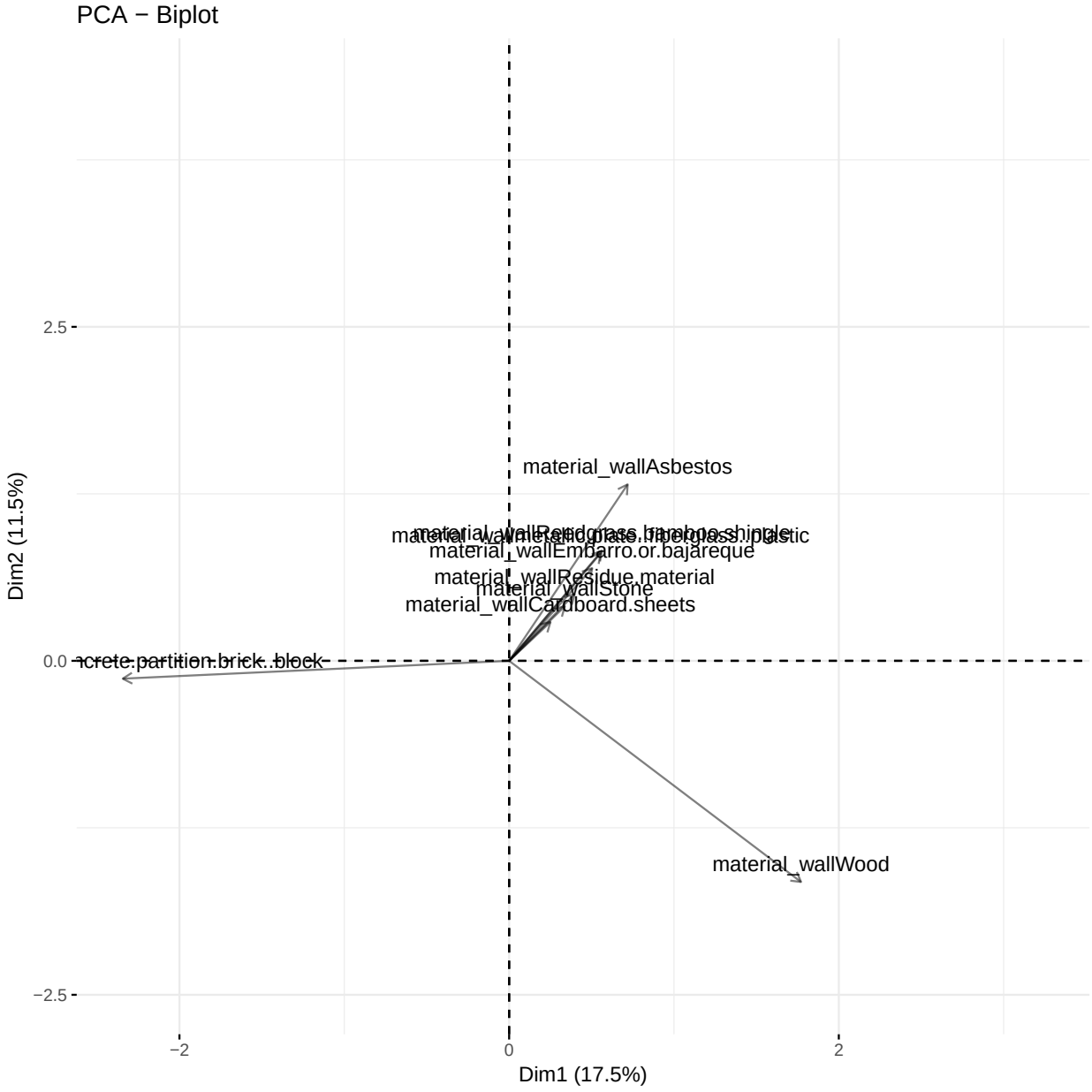


Part 2 - PCA categorical variables

First round

Categorical variable: `material_wall`.

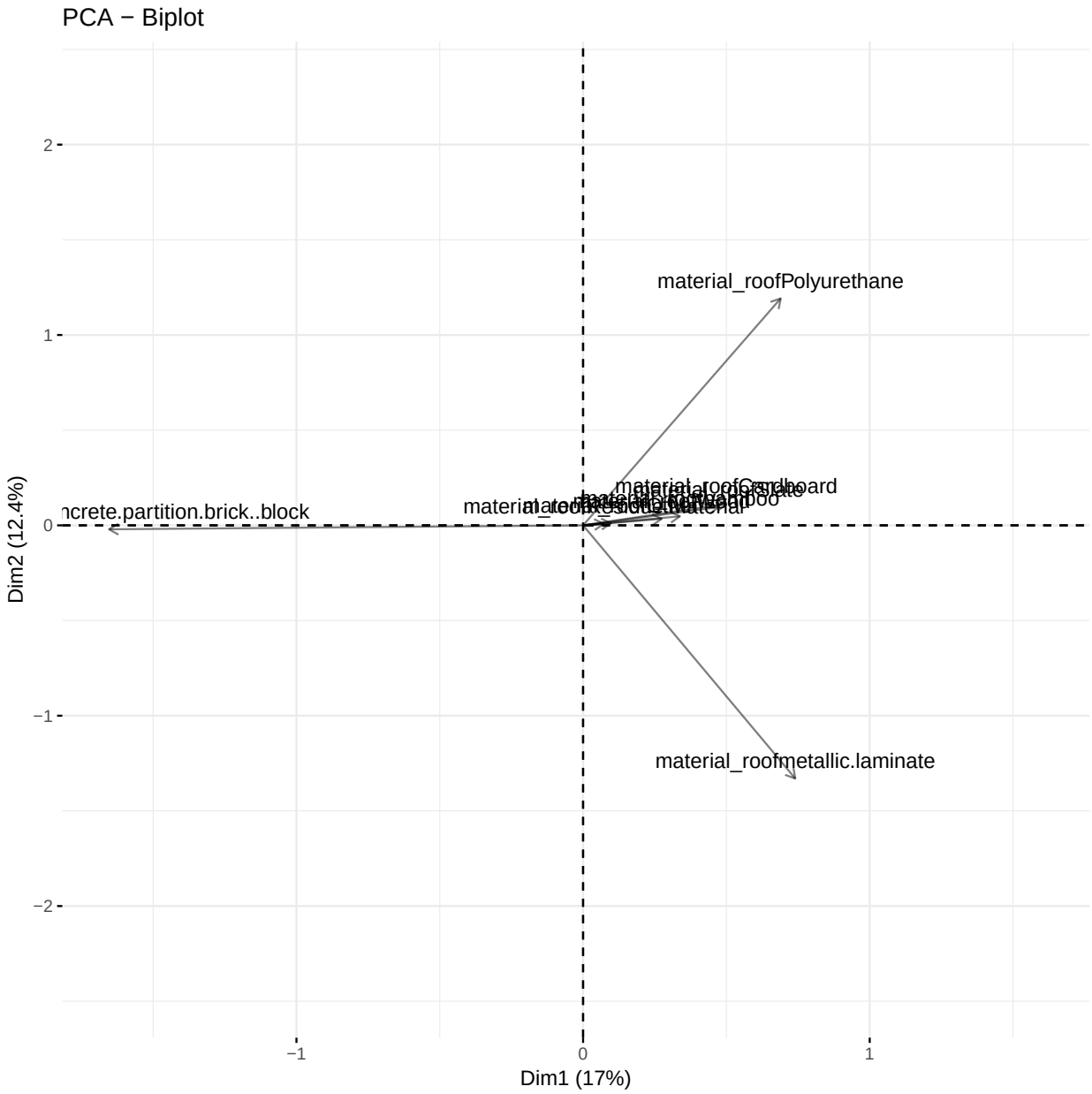
Dummy variables kept: `material_wallAsbestos`, `concrete.partition.brick..block`, `material_wallWood`



Second round

Categorical_variable : `material_roof`.

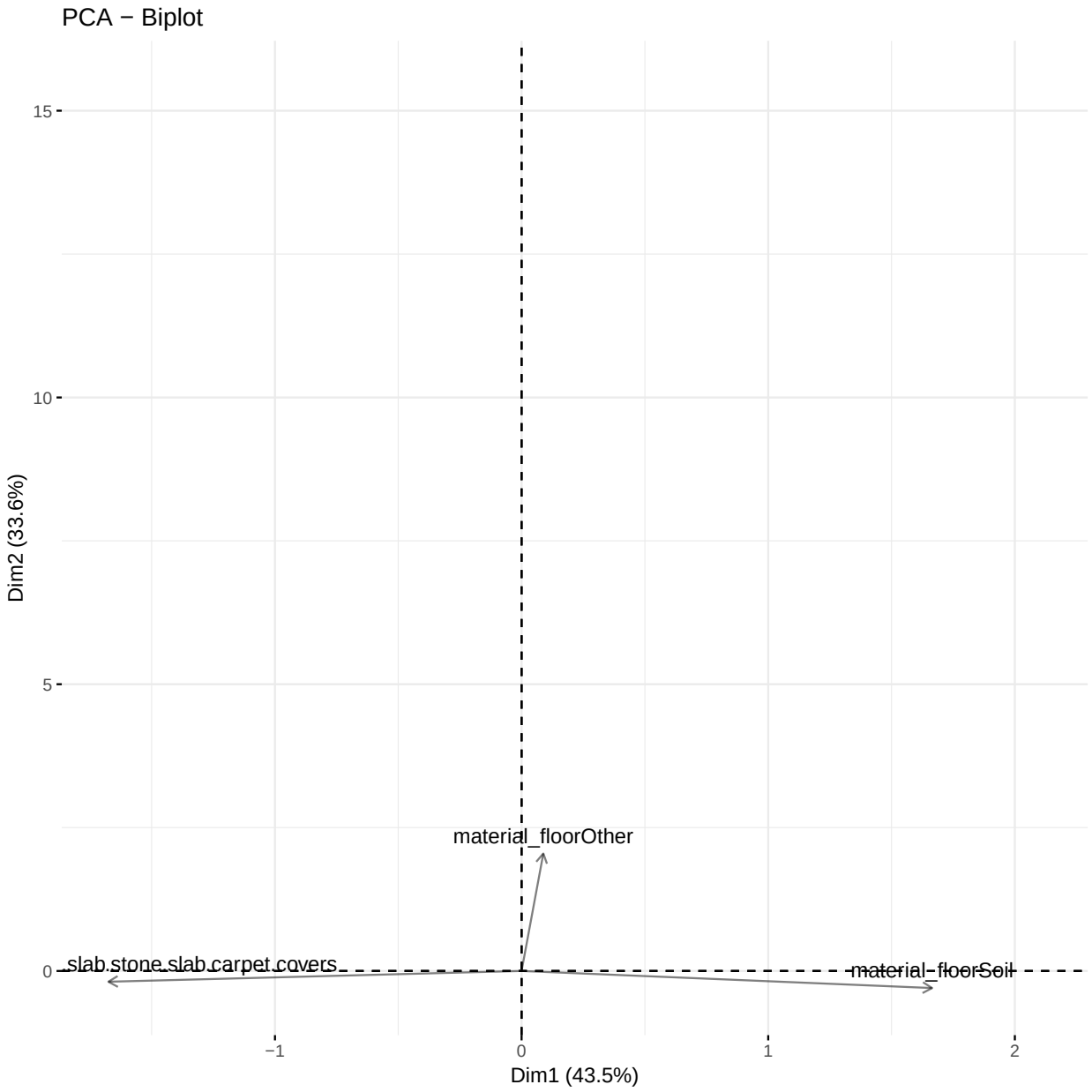
Dummy variables kept: `material_roofConcrete.partition.brick..block`, `material_roofmetallic.la`
`material_roofPolyurethane`, `material_roofCardboard`.



Third round

Categorical variable: `material_floor`.

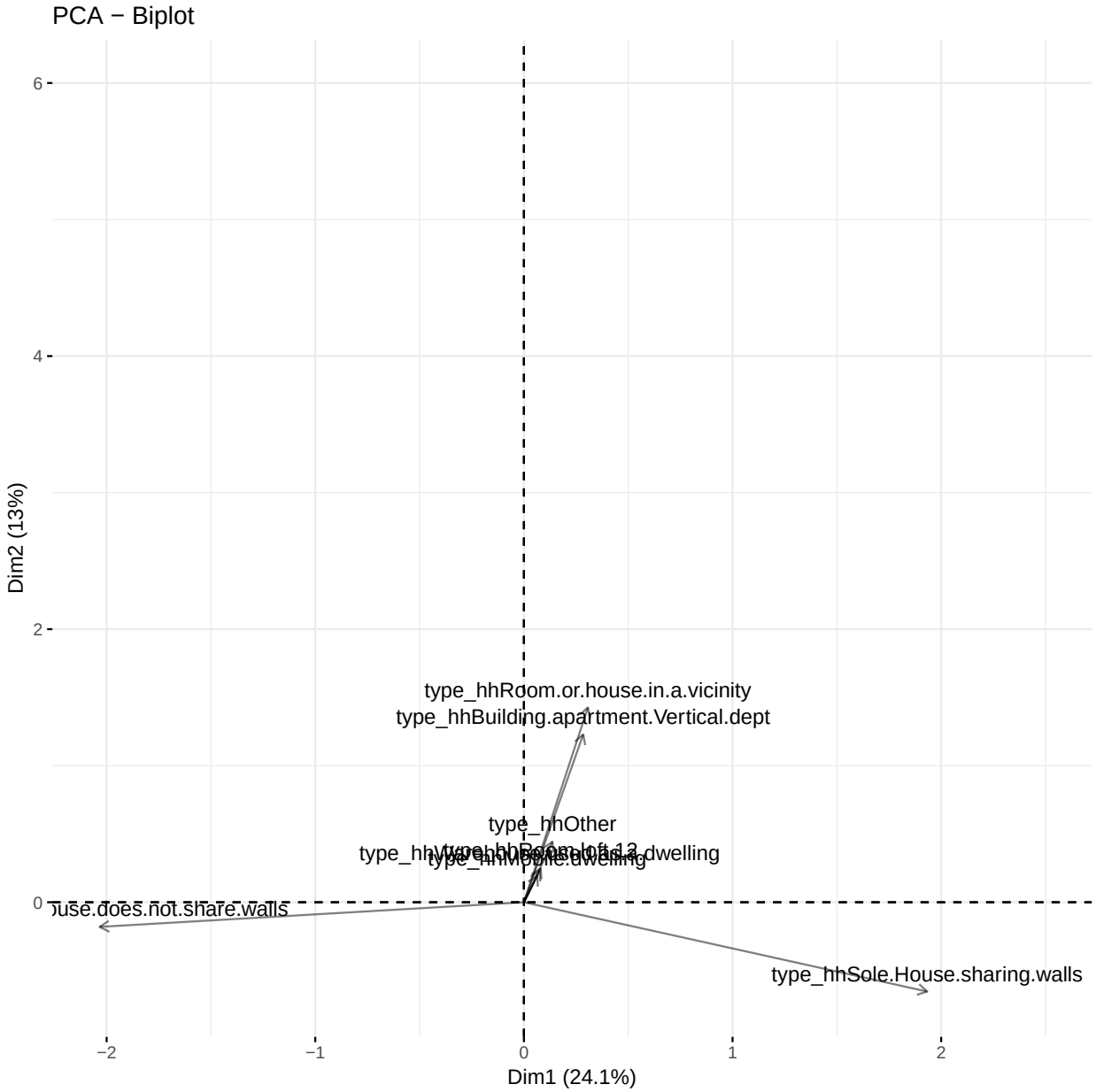
Dummy variables kept: `material_floorSoil`, `material_floorWood.slab.stone.slab.carpet.covers`, `material_floorOther`.



Fourth round

Categorical variable: `type_hh`

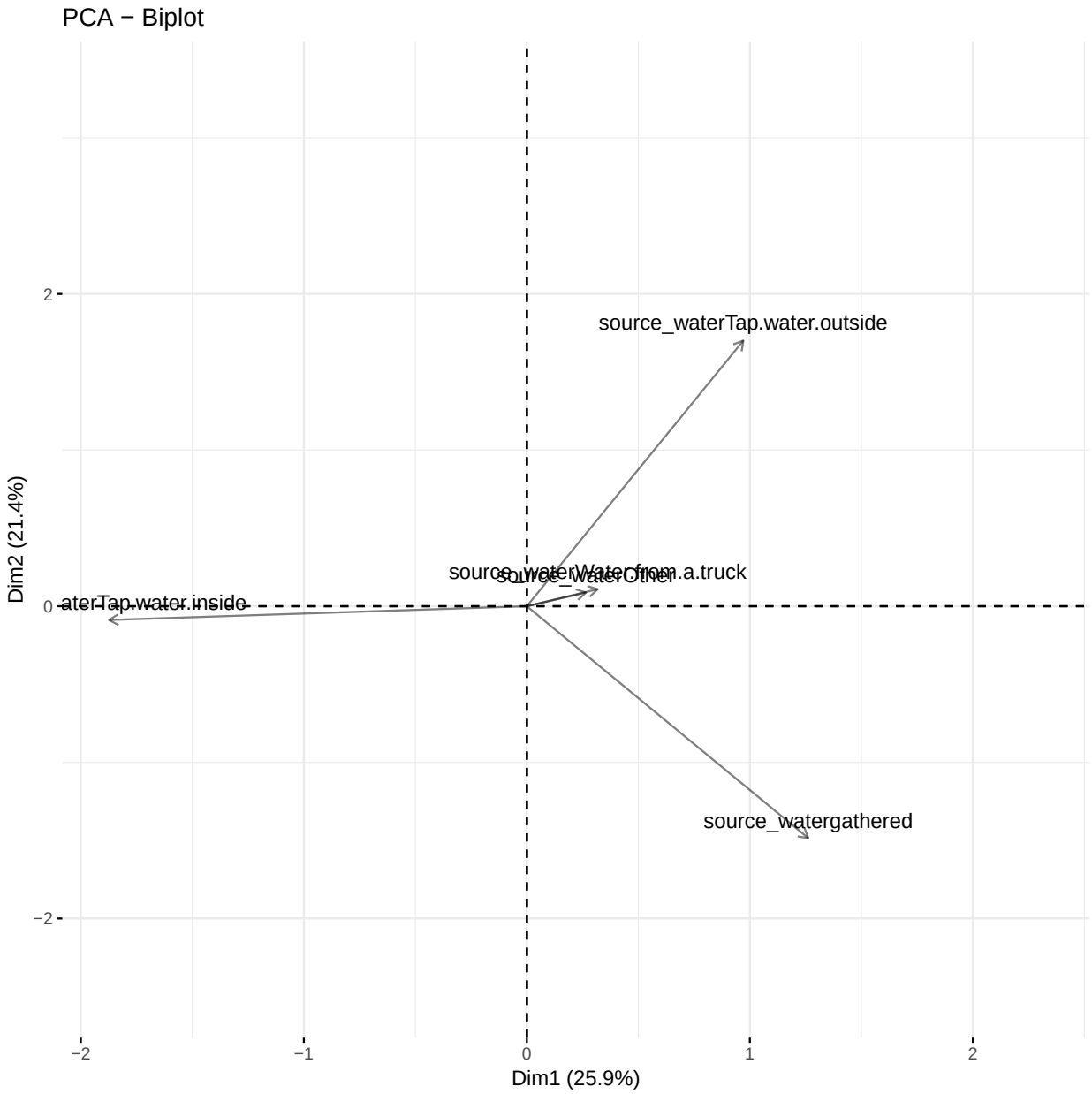
Dummy variables kept: `type_hhSole.House.sharing.walls`, `typehhRoom.or.house.in.a.vicinity`.



Fifth round

categorical variable: `source_water`

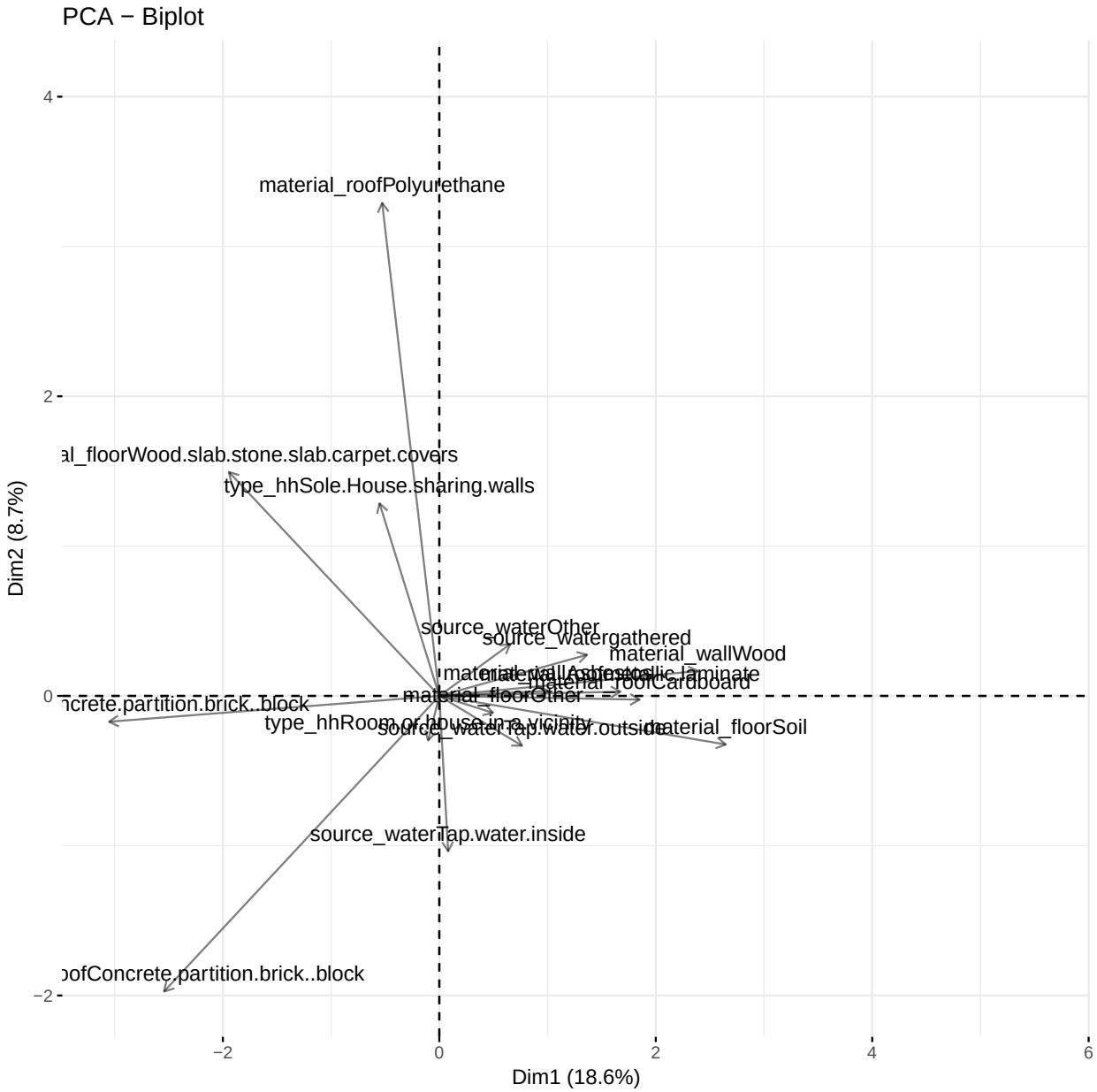
Dummy variables kept: `source_watergathered`, `source_waterOther`, `source_waterTap.water.inside`, `source_waterTap.water.outside`.



Sixth Round

Theoretical category: physical characteristics of the household.

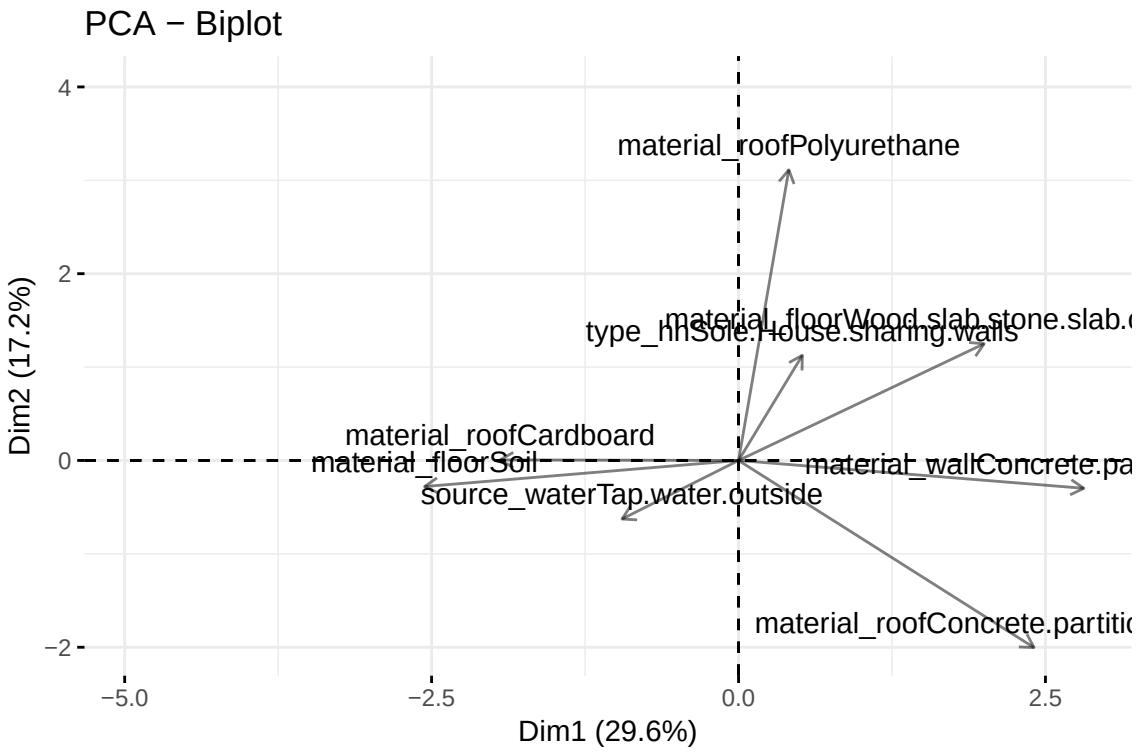
Dummy variables kept: `material_roofPolyurethane`, `material_floorWood.slab.stone.slab.carpet.c` `type_hhSole.House.sharing.walls`, `material_roofConcrete.partition.brick..block`, `material_wallConcrete.partition.brick..block`, `source_waterTap.water.outside` `material_floorSoil`, `material_roofCardboard`.

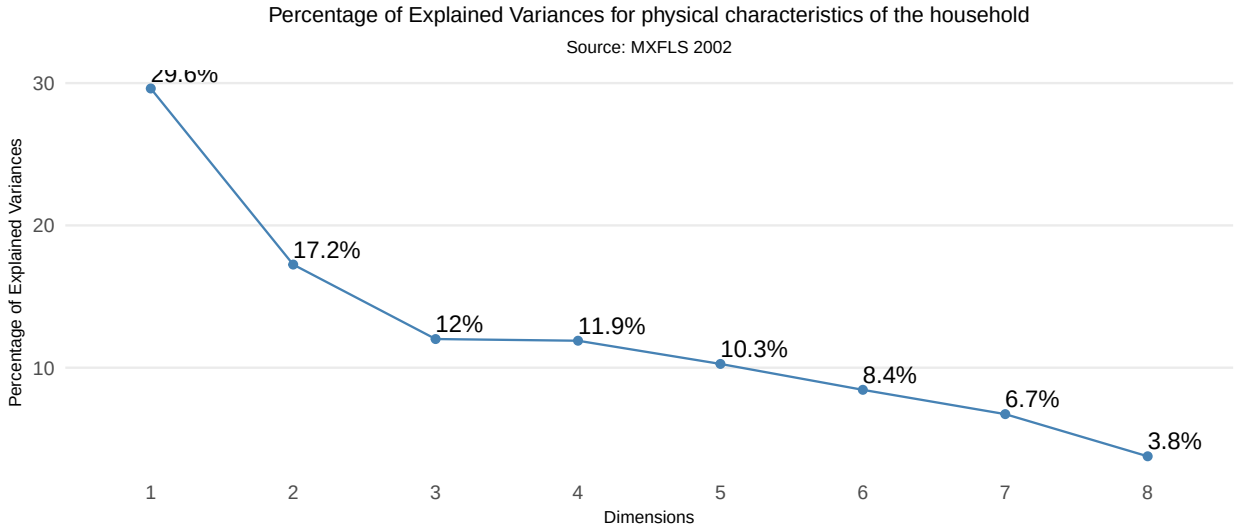


Seventh Round

Theoretical category: physical characteristics of the household.

Dummy variables kept: `material_roofPolyurethane`, `material_floorWood.slab.stone.slab.carpet.c` `material_roofConcrete.partition.brick..block`, `material_wallConcrete.partition.brick..bloc` `material_floorSoil`, `material_roofCardboard`.

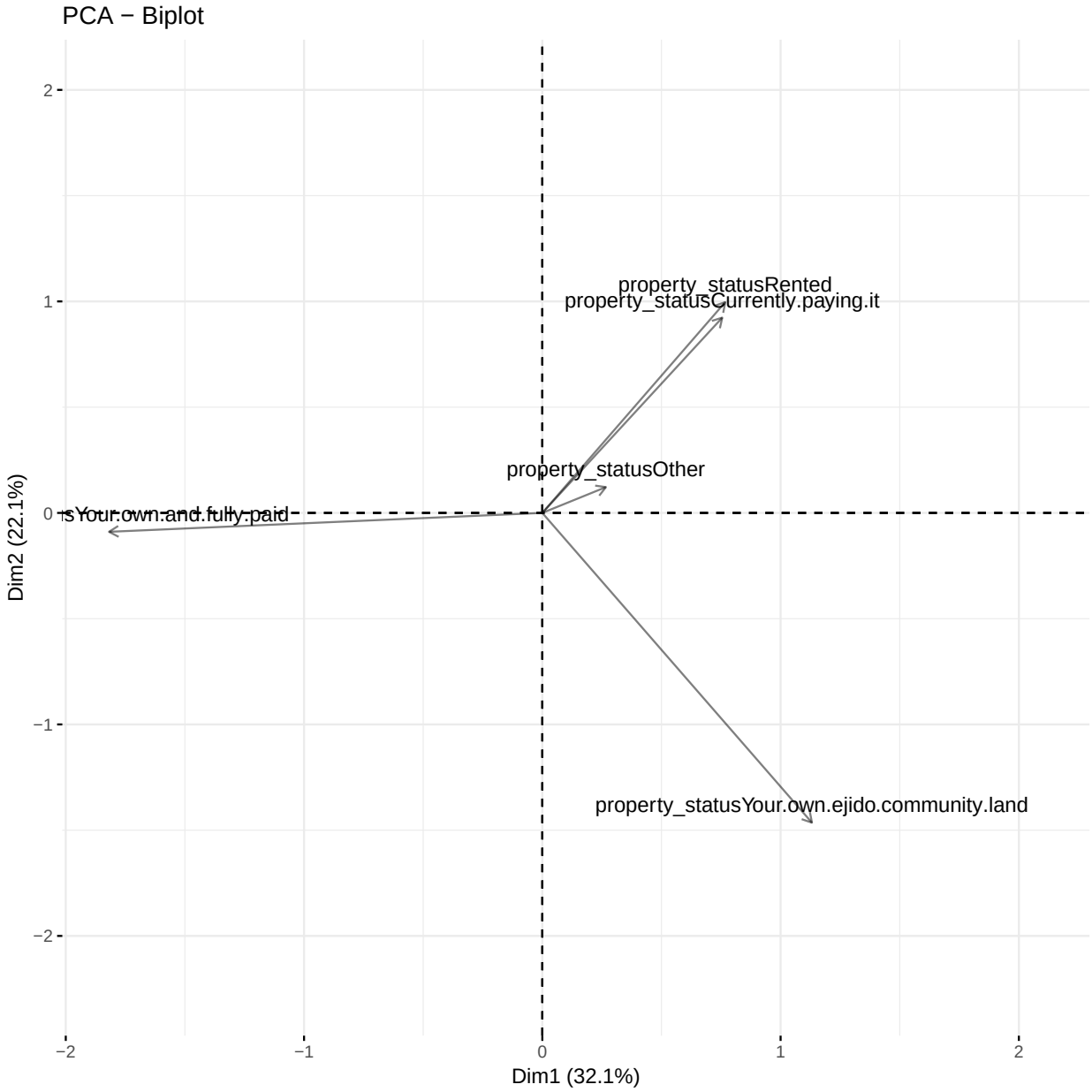




Eighth Round

Categorical variable: `property_status`.

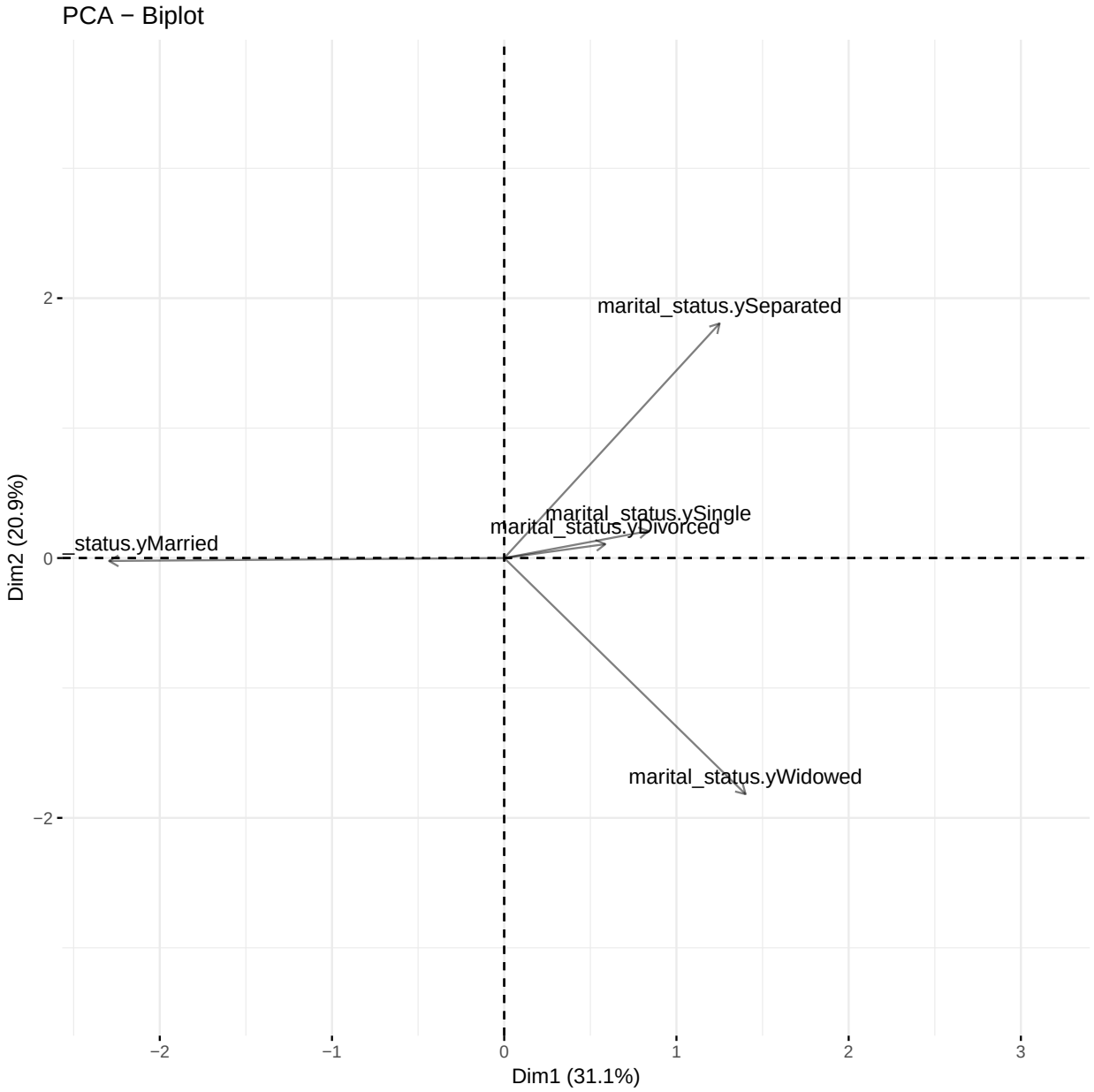
Dummy variables kept: `property_statusOther property_statusRented`, `property_statusYour.own.ar`
`property_statusYour.own.ejido.community.land`.



Ninth Round

Categorical variable: `marital_status`

Dummy variables kept: `marital_status.yMarried`, `marital_status.ySingle`, `marital_status.yWidowed`.



Part 3 - Supervised Learning

CART model

```
## CART
##
## 6336 samples
##   27 predictor
##    2 classes: 'Attend', 'Not_Attend'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1268, 1267, 1267, 1267, 1267
## Resampling results across tuning parameters:
##
##   cp     ROC        Sens       Spec
##   5e-06  0.6481418  0.9805820  0.0672043
##   5e-04  0.6481418  0.9805820  0.0672043
##   1e-02  0.5527893  0.9952736  0.0155914
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was cp = 5e-04.
```

Random forest

```
## Random Forest
##
## 6336 samples
##   27 predictor
##    2 classes: 'Attend', 'Not_Attend'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1268, 1267, 1267, 1267, 1267
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    3    0.7724645  0.9998882  0.0006048387
##    5    0.7728342  0.9995076  0.0034946237
##   10    0.7678904  0.9980119  0.0118615591
##   15    0.7618246  0.9970059  0.0209341398
##
## Tuning parameter 'splitrule' was held constant at a value of gini
##
## Tuning parameter 'min.node.size' was held constant at a value of 10
## ROC was used to select the optimal model using the largest value.
```

```
## The final values used for the model were mtry = 5, splitrule = gini
##  and min.node.size = 10.
```

Out of sample prediction CART model

```
## Confusion Matrix and Statistics
##
##
## pred_CART    Attend Not_Attend
##   Attend       1535         91
##   Not_Attend     20          8
##
##                Accuracy : 0.9329
##                  95% CI : (0.9197, 0.9445)
##     No Information Rate : 0.9401
##     P-Value [Acc > NIR] : 0.9008
##
##                   Kappa : 0.1023
##
##  Mcnemar's Test P-Value : 3.051e-11
##
##             Sensitivity : 0.98714
##             Specificity : 0.08081
##          Pos Pred Value : 0.94403
##          Neg Pred Value : 0.28571
##              Prevalence : 0.94015
##          Detection Rate : 0.92805
##    Detection Prevalence : 0.98307
##       Balanced Accuracy : 0.53397
##
##        'Positive' Class : Attend
##
```

Out of sample prediction forest model

```
## Confusion Matrix and Statistics
##
##
## pred_forest  Attend Not_Attend
##   Attend       1555         98
##   Not_Attend      0          1
##
##                Accuracy : 0.9407
##                  95% CI : (0.9283, 0.9516)
##     No Information Rate : 0.9401
```

```
##     P-Value [Acc > NIR] : 0.4854
##
##                   Kappa : 0.0188
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0101
##          Pos Pred Value : 0.9407
##          Neg Pred Value : 1.0000
##              Prevalence : 0.9401
##          Detection Rate : 0.9401
##    Detection Prevalence : 0.9994
##       Balanced Accuracy : 0.5051
##
##        'Positive' Class : Attend
##
```

Covariate Importance (CART model):