The Impact of Financial Access on School Attendance in Mexico

Group No. 5

April 29, 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

School attendance has been considered important to economic growth because of its association with productivity. The general idea behind this is that people with higher education have higher income, as evidenced by Ashenfelter (1997). There are many factors that could affect why a child attends or does not attend school. In particular, financial access could affect schooling through access to credit and savings with better conditions. According to Becker (1975), by limiting the incentives and capacity to invest in human capital, credit constraints play an important role in determining aggregate productivity, national income distributions, social mobility, and economic growth and development. In Aportela (1998) it was found that access to financial services increases savings on low-income people. What this could imply is that with better access to financial instruments that allow people to save money, a household head may better prepare for the expenses that will have to be made when the child goes to school. In other words, by having access to these financial services a household head may be able to ask for a credit in order to meet their basic needs instead of putting a child to work.

According to Augsburg, Harmgart and Meghir (2012) there are several channels through which microfinance can impact education. One possibility is that microfinance through the loan and the expanded business activity alleviates liquidity constraints and leads to an expansion of school attendance and a decrease in child labor. With a sufficiently low interest rate and a sufficiently high rise in productivity thanks to schooling, households likely will decide to attend school because the financial market allows them to smooth out their consumption; however, the perception of the value of schooling is fundamental in the decision of attending or not attending school.

There are multiple data sets which will comprise the source of our analysis. The main source for the outcome `attend_school` and the control characteristics are in the Mexican Family Life Survey (ENNVIH). The ENNVIH is a set of multiple books each containing a set of data sets. We used the control book for 2002 (M X F L S - Book C). This book contains several control characteristics at an individual level (i.e. marital status, maximum education grade, etc.) and at a household level (i.e. proximity to water source, type of floor, etc.). The second book of data sets are contained in 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. They include but are not limited to, child employment, child health characteristics and more[1]. The last group of data sets is contained in the Bank Operational Information from the Comision Nacional Bancaria y de Valores (CNBV). This data set contains information about the locations of bank branches in 2002 for Mexico[2].

Methods and Research Design

The first step consisted of creating a unique data set which contained all of the pertinent covariates from the multiple data sets. Extensive use of the `tidyverse` was necessary to recode, reorganize, merge and tidy the data. One issue with the `tidyverse` was the difficulty in running loops through the `mutate` function and creating functions which included `mutate`; however, the issue had a solution within the same package. Once the unique data set was constructed by merging all of the data sets in M X F L S - Book C & M X F L S - Book V, it contained over 90 covariates —even after removing redundant variables, variables with excessive percentage of missing values and containing little useful information— we used Principal Component Analysis (PCA) to reduce the number of covariates substantively; the `FactoMineR` and `factoextra` packages were used for this purpose. This was a labour intensive work and one suggestion could be having as output the suggested variables to drop.

Even after running PCA there were still a considerable amount of covariates. For this purpose we used Supervised Machine Learning techniques to explore which covariates were the most important in predicting `attend_school`. The idea behind this was to be able to eliminate covariates which didn't contribute much to the prediction of school attendance; by doing so,

[1] http://www.ennvih-mxfls.org/english/ennvih-1.html

[2] http://portafoliodeinformacion.cnbv.gob.mx/bm1/paginas/infoper.aspx

it would provide evidence (minimal as it may be) that the included covariates when running the regression to ascertain the effects of bank access on school attendance were enough to eliminate or minimize bias in the regression coefficients. We ran two models in order to do this, a random forest —modifying the number of covariates included as a tuning parameter— and a classification and Regression Tree (CART) —modifying the complexity parameter as a tuning parameter— and chose the one with highest specificity metric to obtain the list of covariates which were most important in predicting `attend_school`.

The last step in our research design was supposed to be a regression analysis to estimate the impact of bank proximity or bank access on school attendance. The problem was that the data set contained little variation in the outcome variable `attend_school`; over 93% of children over the age of 5 attended school. What this meant was that in addition to all of the usual limitations that regression analyses contain, our results would have likely been non-informative[3]. With this problem in mind, we decided to explore the data set and make visualizations for potential future research; we included a density map of Mexico by State showing the ratio of the numer of persons to the number of bank branches per state.

Tools

Constructing Data

Manipulating the data and preparing it for analysis required a great deal of work; we briefly 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.

We started by loading the control book for 2002 (M X F L S - Book C). 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

[3]A potential research design for the future could contain manipulations to the outcome variable to account for social desirability bias so that the data more accurately reflects reality.

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 construcction 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 [4].

We then proceeded to load individual level data for children by loading the book V for 2002 (M X F L S - Book V). Within the book there are six pertinent data sets to be used [5]:

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.

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).

[4] For further details as to the cleansing of the data check Appendix - Part 1 - M X F L S Book C.

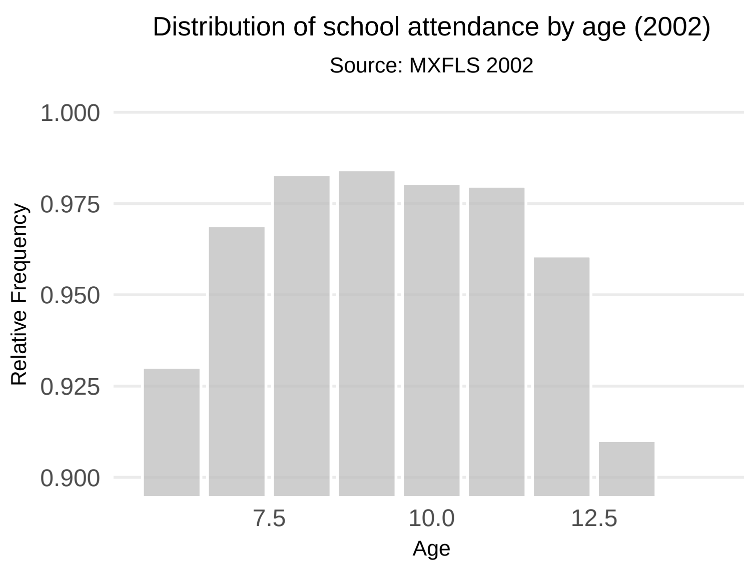
[5] For further details regarding the wrangling of these data sets check Appendix - Part 1 - M X F L S - Book V.

Lastly, 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 about 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 names as opposed to numbers so another data set called `Relacion_de_municipios_de_Mexico.xlsx`, which contains the relationship between municipalities names and their official corresponding numbers, was used[6].

Exploratory Analysis

School Attendance Distribution

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



It is important to notice the scale on the y-axis, for every age group children are attending

[6]The CNBV data set contains an exhaustive list of bank branches, therefore municipalities not included in this data set do not have any bank branches; what this implies is that the missing values for the bank dummy when merging the ENNVIH data set and the CNBV data set are really values of zero.

school at rates of over 90 percent —this influenced our goodness of fit measure in the supervised learning section—. This graph is very important because it shows that there is likely social desirability bias. Therefore, the estimated impact of access to banks would be very unreliable even if the results were stratified by age to allow for the effect to be different by different groups of age. [7] This is the main reason why future research could focus on correcting for social desirability bias in order to run proper regression models; however, this task falls out of the scope of this project.

Principal Component Analysis (PCA)

Since we have many covariates in our data set, we looked at the correlation matrix to detect any potential problems with the variables [8]. 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. The correlation coefficients in the matrix are fairly small; 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[9].

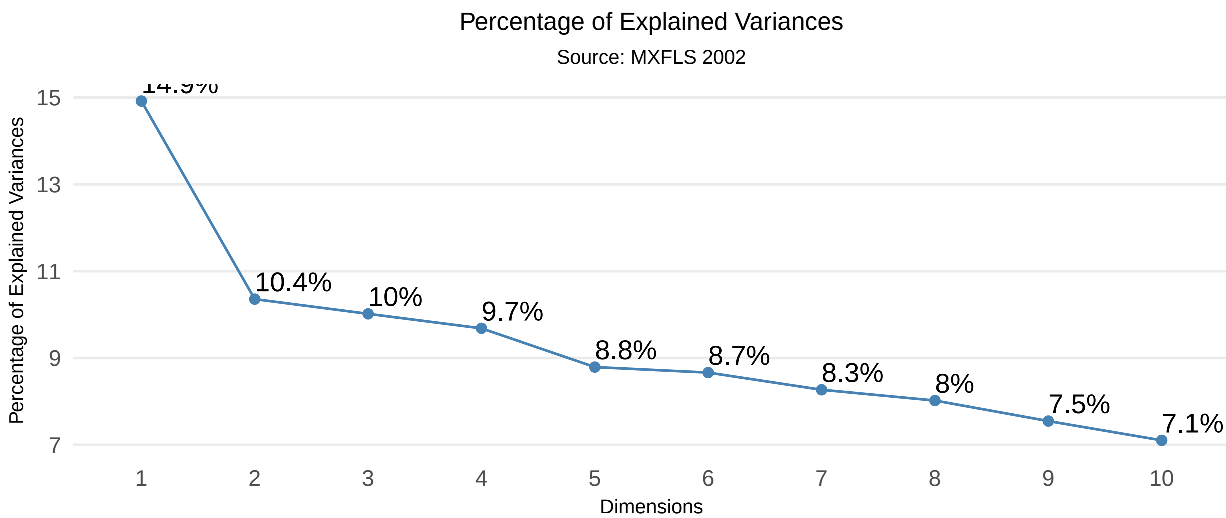
We proceeded to filter out these variables (keeping a single one for reference) to condense our

[7] To illustrate how this could have an impact on the aggregated or average effect (average across groups of age to be specific) refer to Appendix - Part 2 - Exploratory Analysis - Sample Age Distribution.

[8] To look at the correlation plot refer to Appendix - Part 2 - Correlation Matrix

[9] The results can be observed in Appendix - Part 3 - PCA - Continuous Variables

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 for continuous variables as much as possible.

We repeated the same principal component analysis (PCA) for categorical variables[10]. 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 variance. The same process was repeated for categorical variables `marriage_status` (i.e. single, married,

[10] Refer to Appendix - Part 3 - Categorical Variables

divorced, etc.) and property_status (i.e. rent, own, communal land, etc.). Lastly, from over 90 covariates in the totality of the data set we managed to reduce the number to 30 through this process.

Supervised Learning

Once the covariates were significantly reduced we proceeded to build 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.

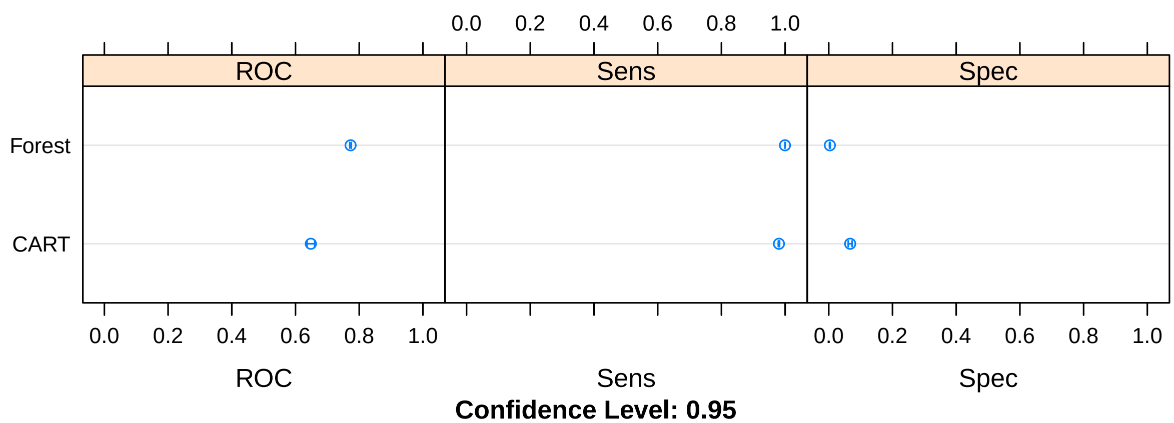
We then ran a CART model and tuned the complexity parameter until we found results that converged to a maximum level of specificity[11]. The second model we ran was a random forest, the packages ranger and e1071 were needed for this purpose. The tuning parameter was mtry which states how many predictors the model can take[12]. More details are provided in following sections, but in general we reduced 30 covariates to a total of 9.

[11] For more details check Appendix - Part 4 - CART Model.

[12] For details check Appendix - Part 4 - Forest Model.

Results

We started by examing the supervised learning models to assess which one had the highest specificity level. Even though specificity was low (which is to be expected given the high levels of school attendance) we can observe from the following graph that the CART model had the highest level[13].

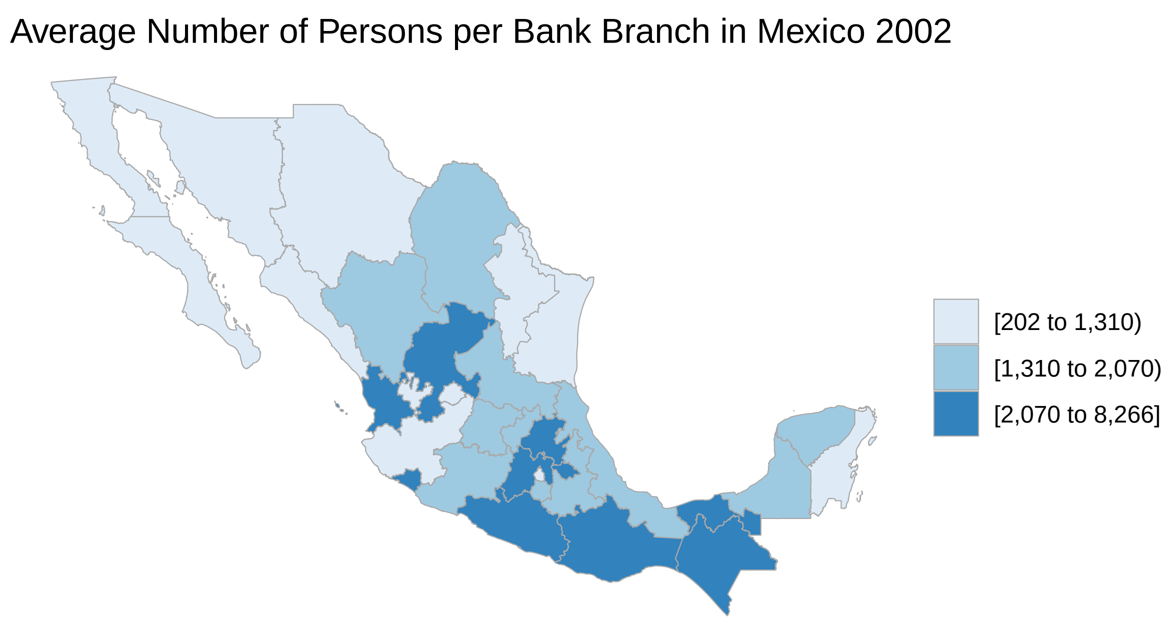


We chose the CART model to obtain information about the predictive accuracy for the variable `attend_school` of each covariate. The top 6 most important covariates in predicting school attendance provided intuitive results[14];The covariates were `age` age of the child at time of interview, `hours_sleep` 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. All of these could plausibly be factors in school attendance and they seem to conform to intuition.

Since running regressions would have produced unreliable results due to the previously mentioned factors, the last step we did was constructing a map of Mexico by State to show the ratio of population per state to the number of bank branches. The idea is to provide a basic tool and potentially think about the number of bank branches and their relationships

[13]This was corroborated by testing the model in out of sample data Appendix - Part 4
[14]For more details see Appendix - Part 4 - Covariate Importance (CART model)

with other variables since it is fair to say that banks do not randomly open bank branches. To do this the package `mxmaps`[15] and the package `sf` were necessary. In addition to this, we merged the total population by state from a file called `Tabulado.csv`[16].



As a general observation, the southern states in Mexico are the most poor, and from the previous map we can observe that the ratio of the number of persons to bank branches is the highest in these states. One could conjecture that banks open branches where there is higher GDP per capita, regardless of total GDP per locality. This could potentially form the base for future research.

[15]The package is only available through Github, so the ‘devtools‘ was additionally required to load it.
[16]Available in the Data folder and from Mexico’s Statistics Office http://www.beta.inegi.org.mx/app/tabulados/default.html?nc=mdemo02

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

- Aportela, Fernando. (1998) Effects of Financial Access on Savings by Low-Income People, Mimeo, MIT.

- Ashenfelter, Orley and Cecilia Rouse (1997). Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins, Quarterly Journal of Economics, Vol. 113, p. 253-284.

- Augsburg, Britta, Ralph De Haas, Heike Harmgart, and Costas Meghir (2013) Microfinance, Poverty and Education. National Bureau of Economic Research, Inc NBER Working Papers 18538..

- Becker, Gary (1975). Human Capital, 2nd ed., Columbia University Press, New York, NY, 1975.

- 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

- Diego Valle-Jones (2018). mxmaps: Create Maps of Mexico. https://www.diegovalle.net/mxmaps/, https://github.com/diegovalle/mxmaps.

- 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

- Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, https://journal.r-project.org/archive/2018/RJ-2018-009/

- 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 - Data Wrangling

M X F L S - Book C

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.

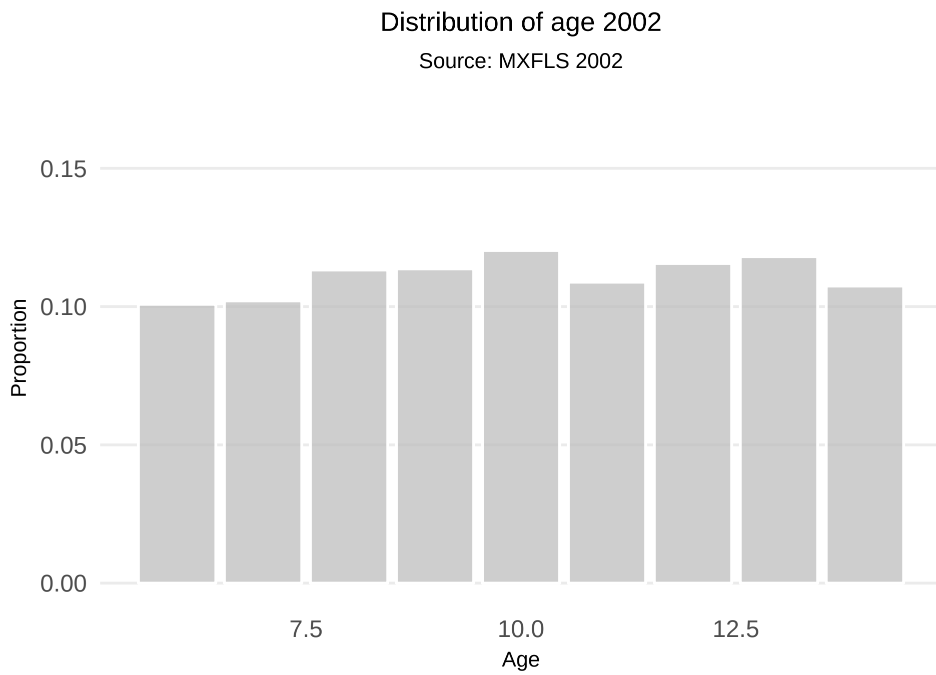
M X F L S - Book V

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.

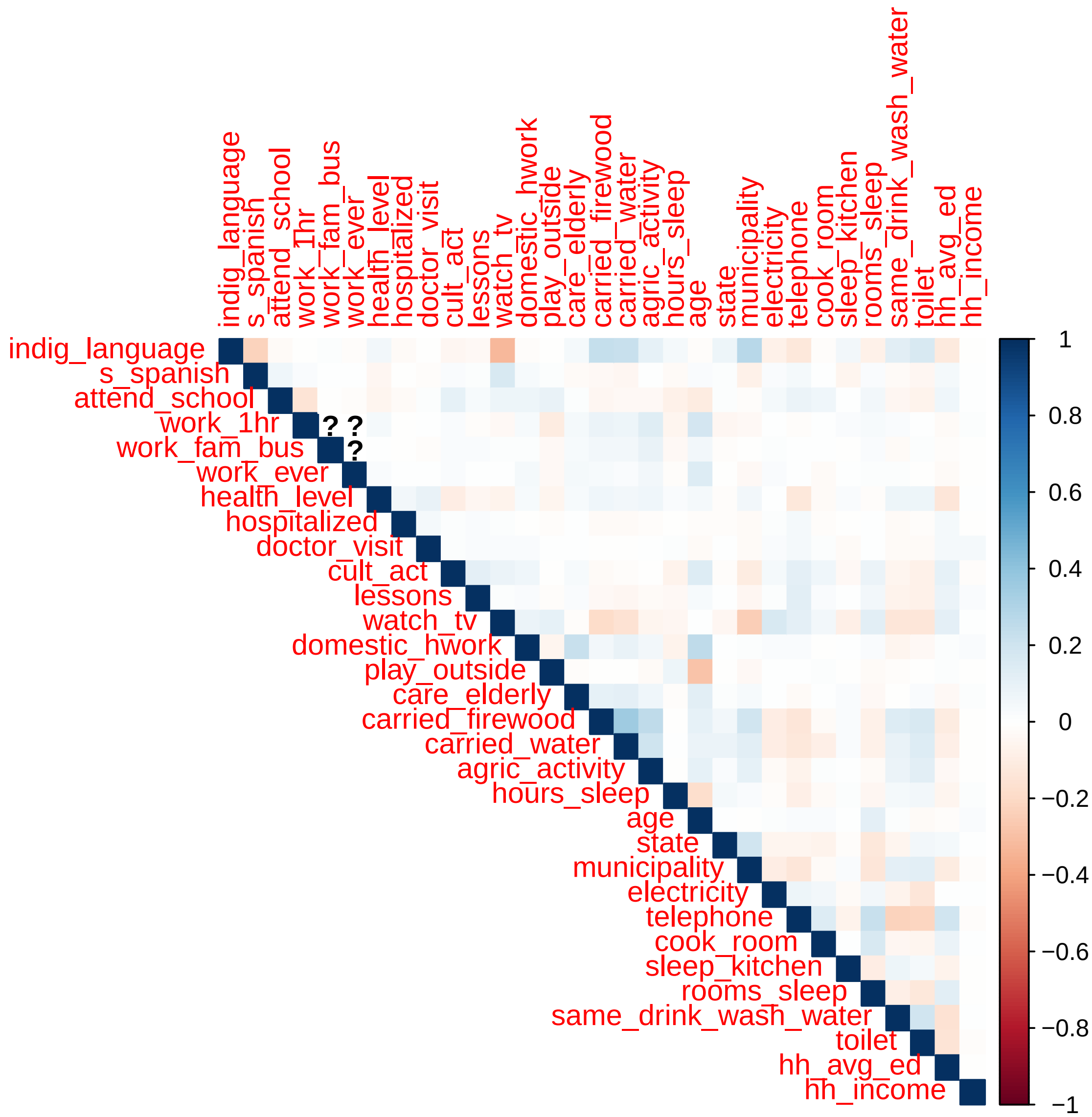
Categorical variables were also transformed to show the actual string category they represent as opposed to numeric value. The last step consisted of recoding dummy variables to have a value of one if condition was true and zero otherwise.

Part 2 - Exploratory Analysis

Sample Age Distribution



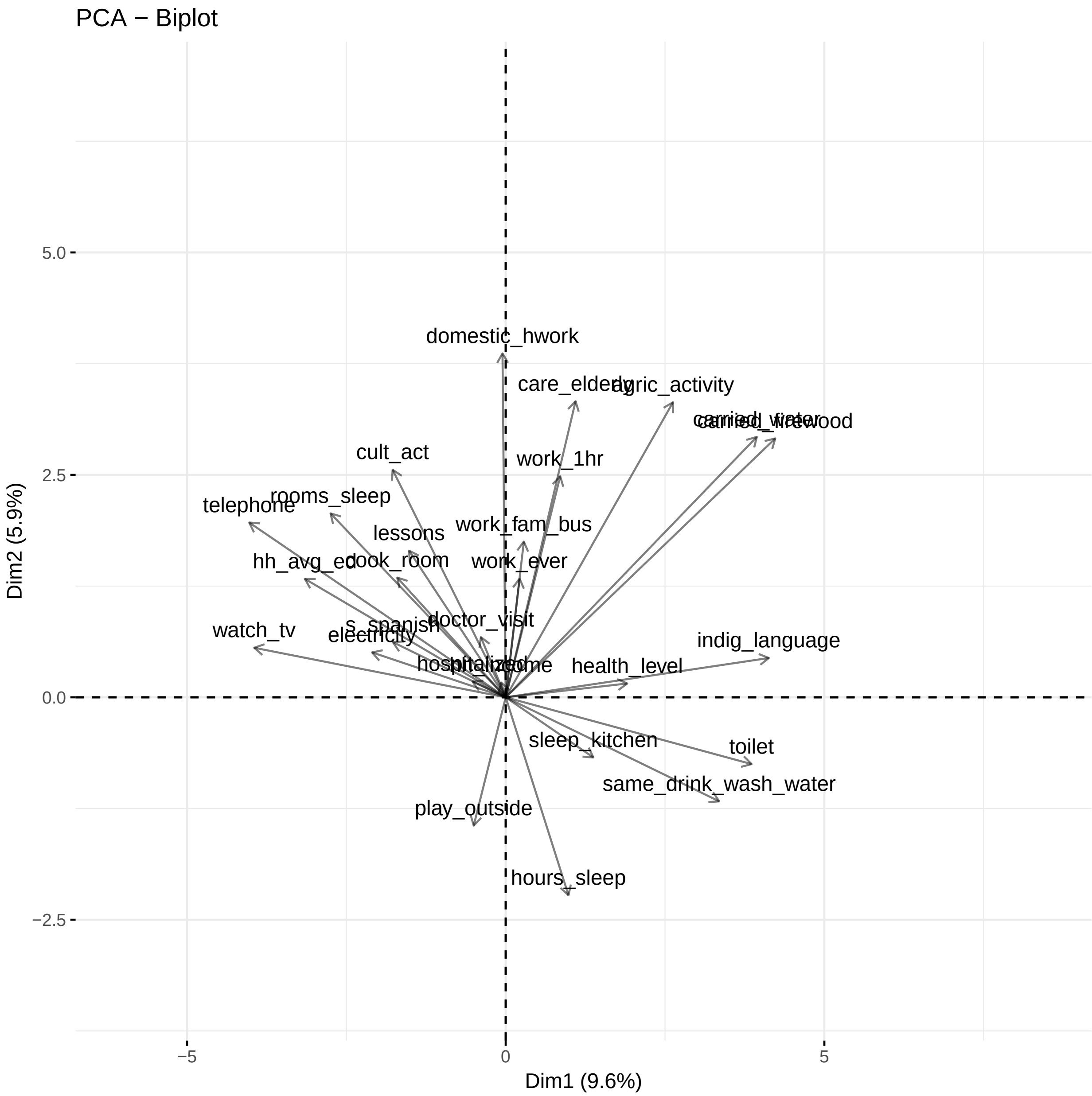
Correlation Matrix



Part 3 - PCA

Continuous Variables

Continuous Variables case:



Some very interesting conclusions from the previous graph include:

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.

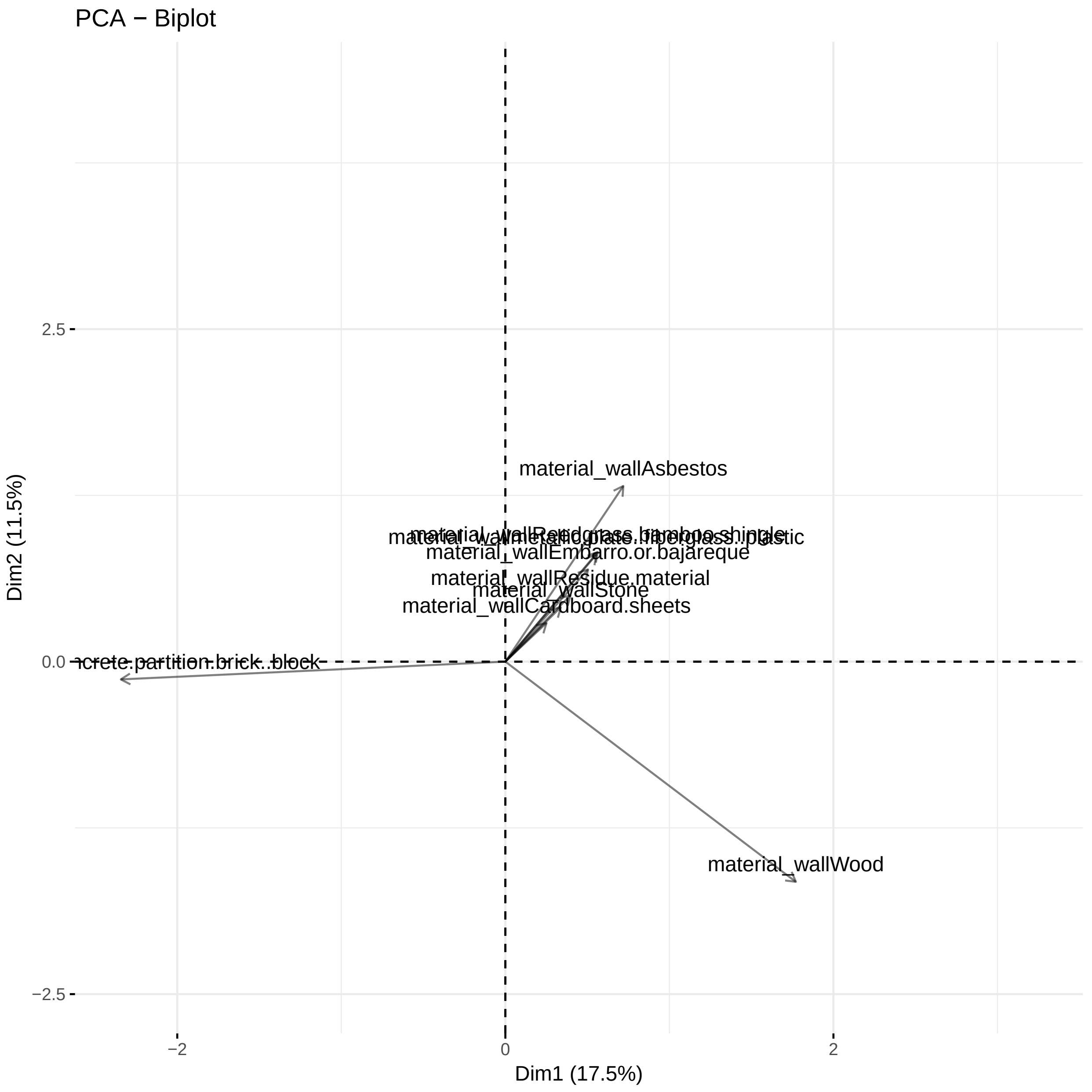
7) `carried_firewood` and `carried_water` seem to be measuring the same thing, carrying a production input or something along these lines.

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.

###Categorical Variables ####First round Ca

Categorical variable: `material_wall`.

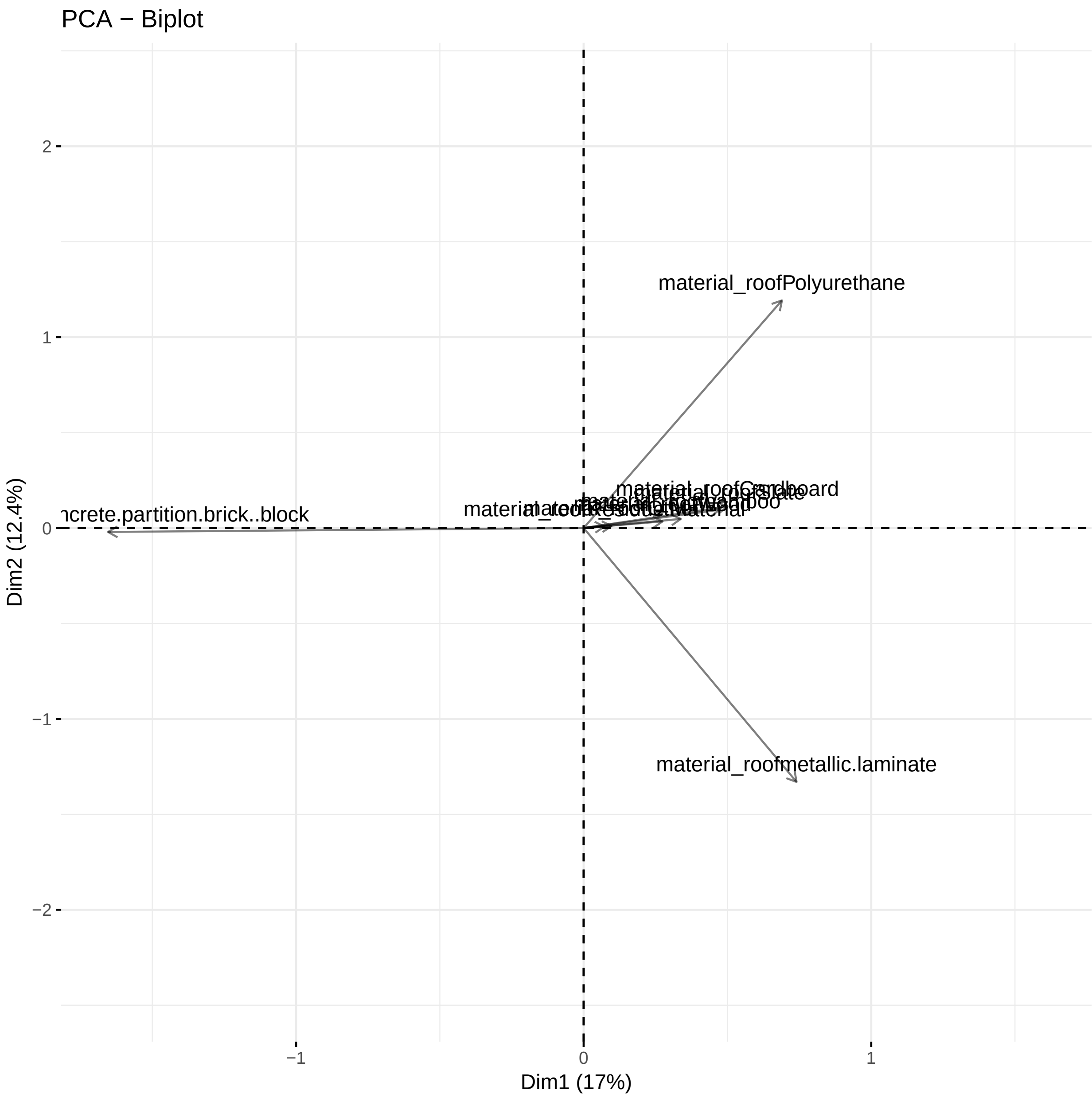
Dummy variables kept: `material_wallAsbestos`, `concrete.partition.brick..block`, `material_wallWood`



Second round

Categorical_variable : `material_roof`.

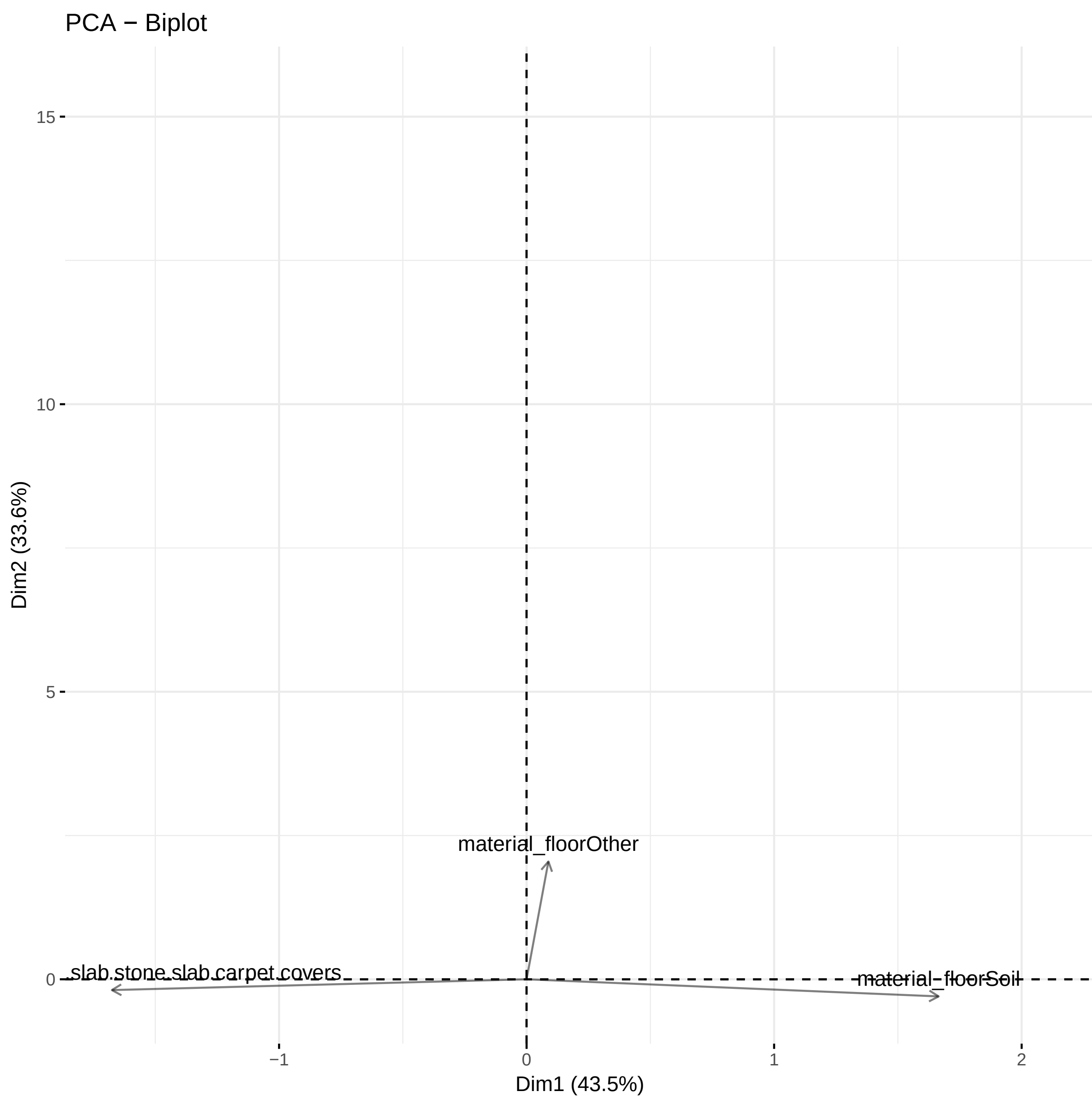
Dummy variables kept: `material_roofConcrete.partition.brick..block`, `material_roofmetallic.la`
`material_roofPolyurethane`, `material_roofCardboard`.



Third round

Categorical variable: `material_floor`.

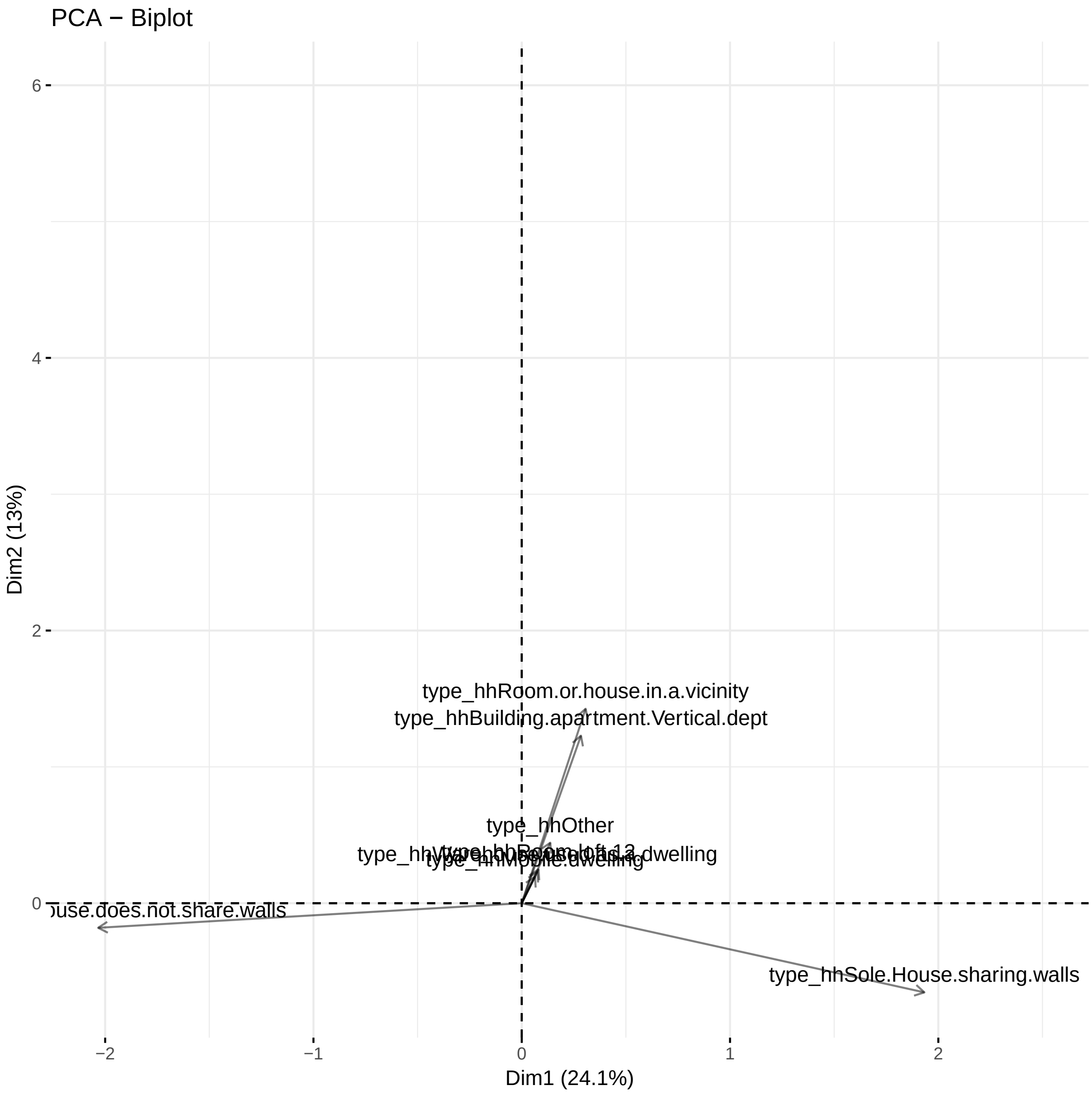
Dummy variables kept: `material_floorSoil`, `material_floorWood.slab.stone.slab.carpet.covers`, `material_floorOther`.



Fourth round

Categorical variable: `type_hh`

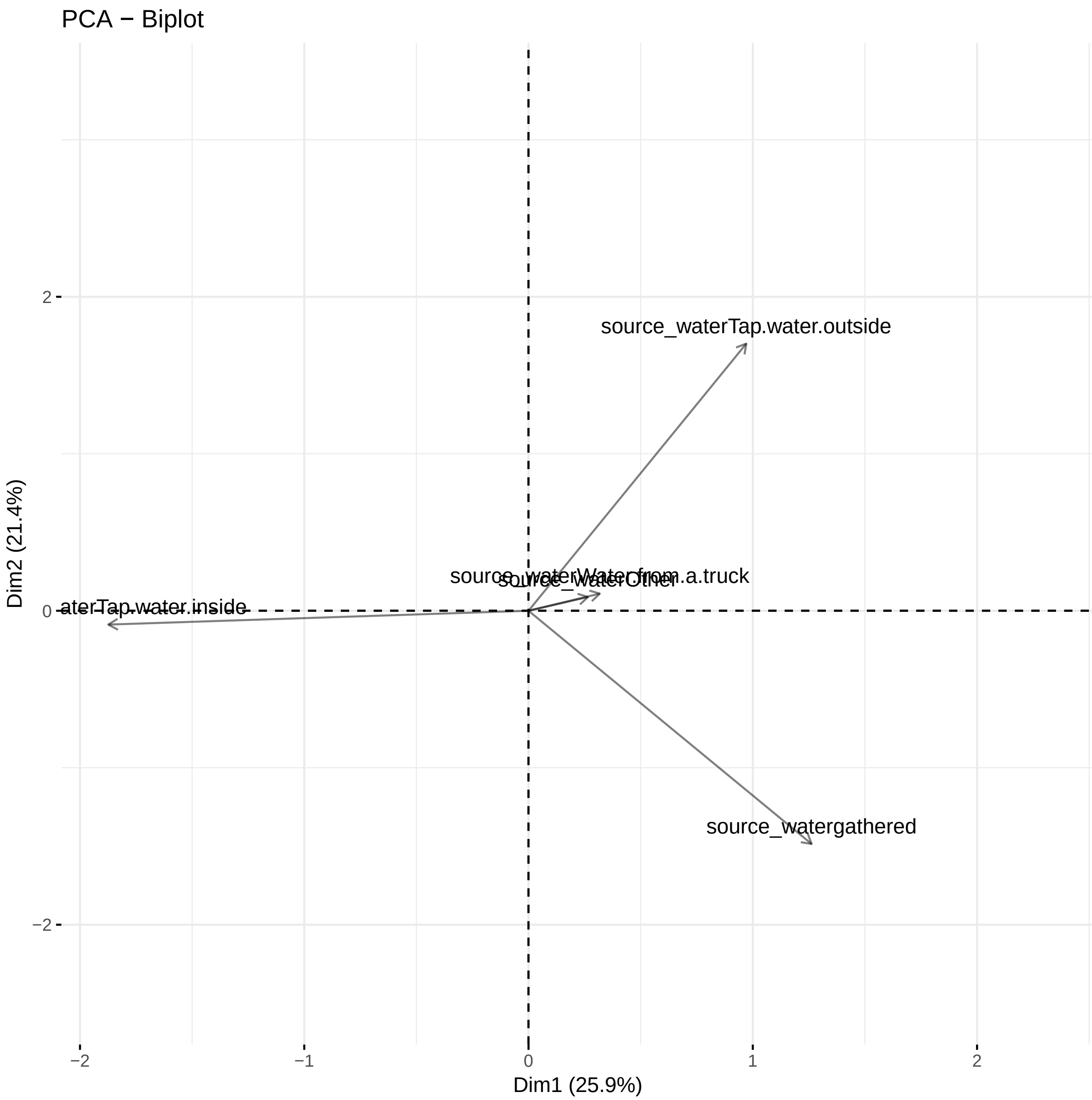
Dummy variables kept: `type_hhSole.House.sharing.walls`, `typehhRoom.or.house.in.a.vicinity`.



Fifth round

categorical variable: `source_water`

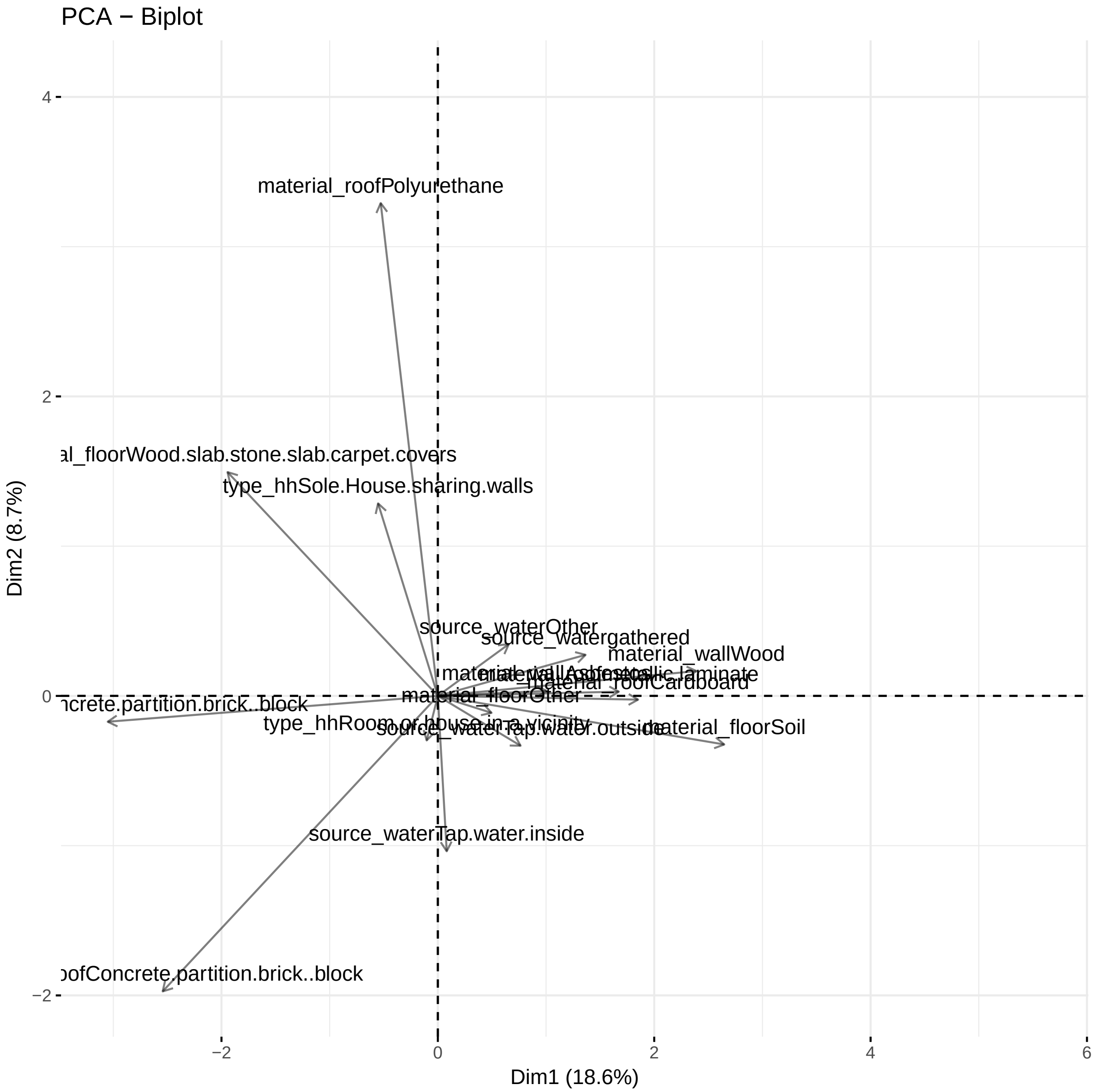
Dummy variables kept: `source_watergathered`, `source_waterOther`, `source_waterTap.water.inside`, `source_waterTap.water.outside`.



Sixth Round

Theoretical category: physical characteristics of the household.

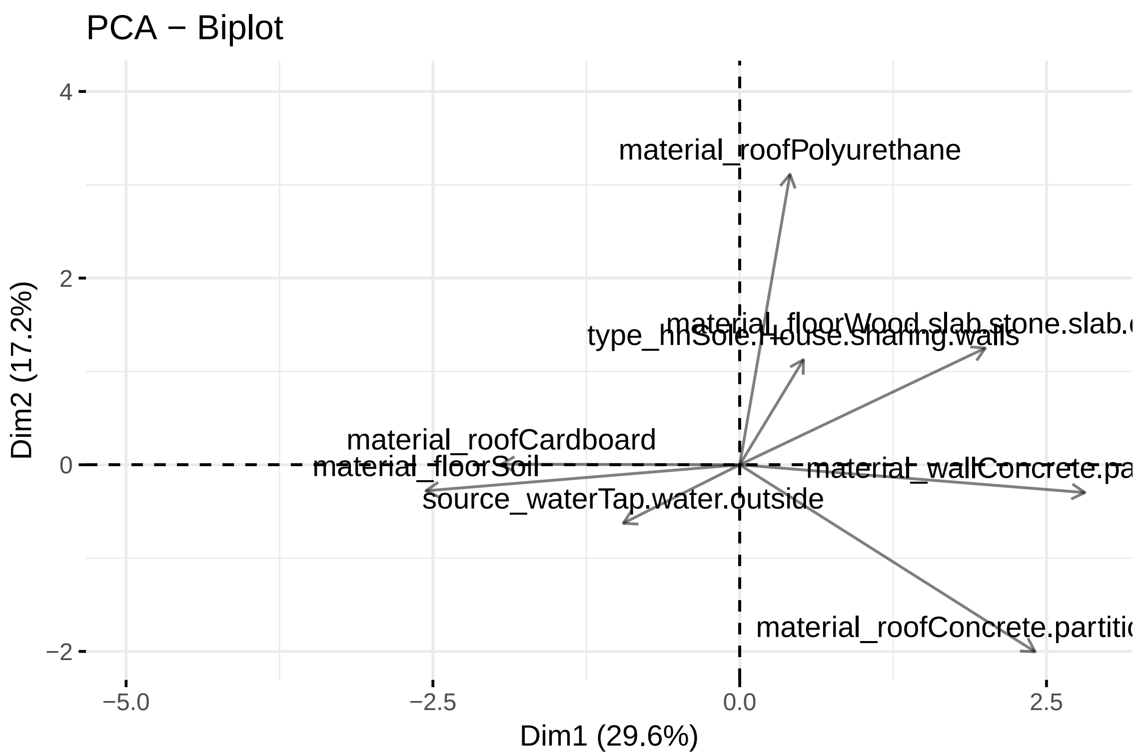
Dummy variables kept: `material_roofPolyurethane`, `material_floorWood.slab.stone.slab.carpet.c` `type_hhSole.House.sharing.walls`, `material_roofConcrete.partition.brick..block`, `material_wallConcrete.partition.brick..block`, `source_waterTap.water.outside` `material_floorSoil`, `material_roofCardboard`.

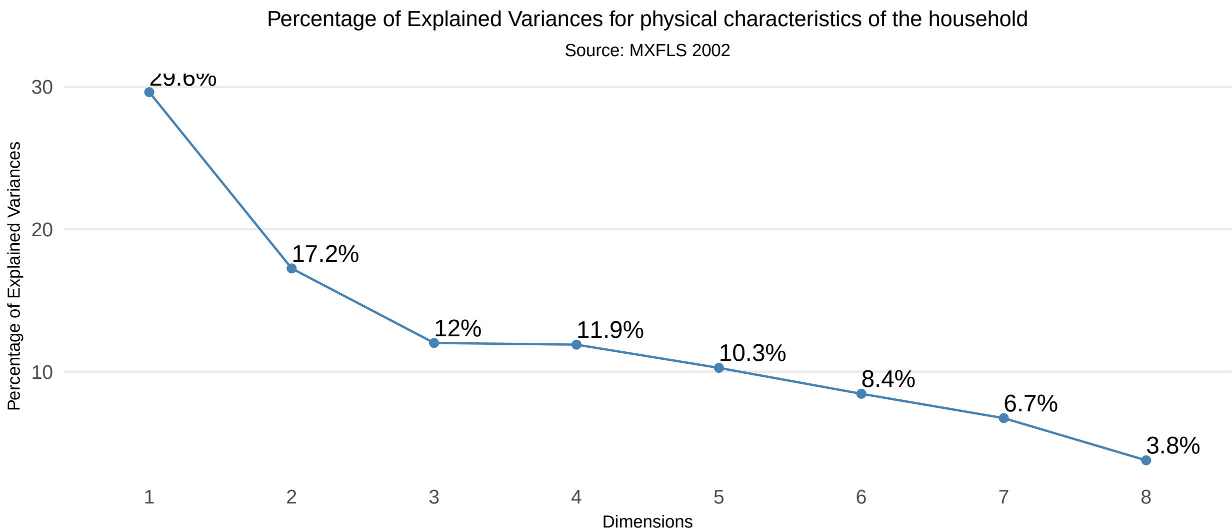


Seventh Round

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`.



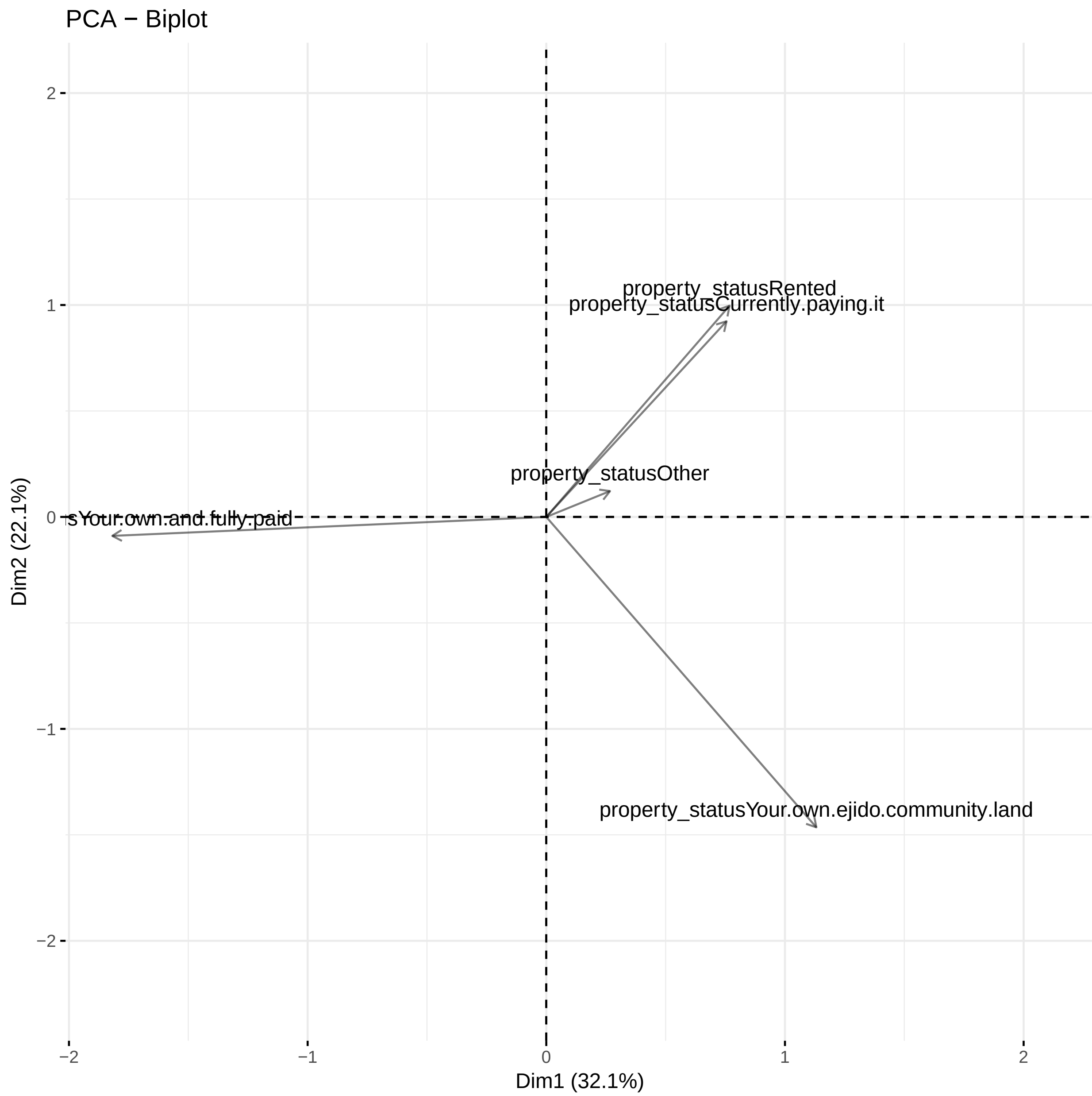


Eighth Round

Categorical variable: `property_status`.

Dummy variables kept: `property_statusOther` `property_statusRented`, `property_statusYour.own.an`

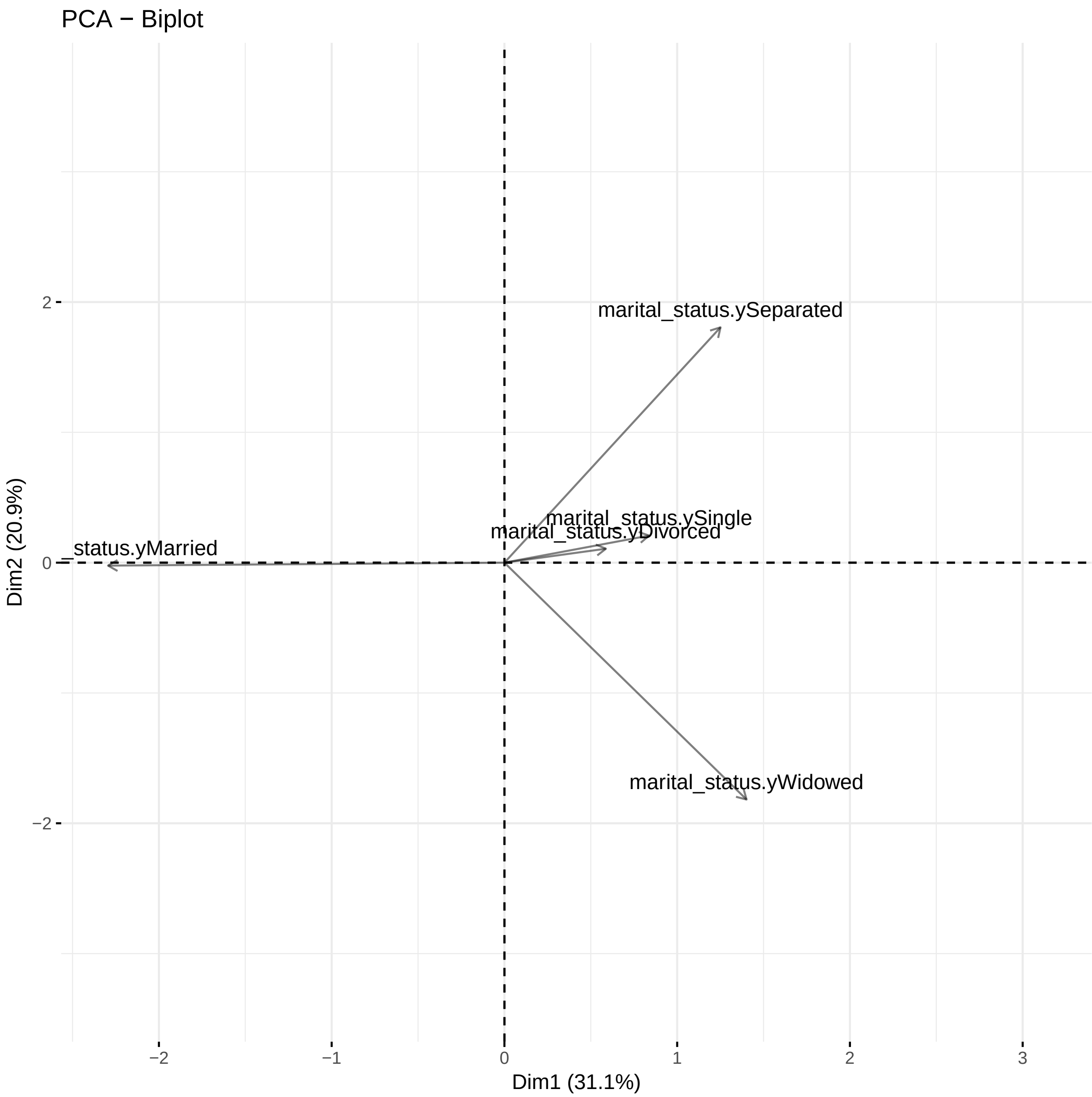
`property_statusYour.own.ejido.community.land`.



Ninth Round

Categorical variable: `marital_status`

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



Part 4 - 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):