Analysis

Creating Modernization Index

Sebastian Nickel

May 04, 2018

TL;DR

This document describes the conststruction of three indices which cover the remoteness and prevailing service infrastructure in villages in Malawi. Two indices are created by the usage of a principal component analysis (PCA), while another is retrieved from a multiple correspondence analysis (MCA).

Variables under considerations

Table 1 shows the variables that will be used to build a villagelevel index. The third and fourth columns show the percentage of missing values for each variable for the two different surveys. The factual health and education (FHE) survey has normally 70% missing values since it only includes a subsample of all villages in the original sample. Hence, the results of the FHE survey are used to fill up or correct the responses from the supervisor survey (SUP).

Also, two distance measures will be included: the average distance to the closest road and the average distance to the district seat.

It is important to note that PCA only works for metric scaled variables, while MCA is used for categorical data. The question q136 needs to be transformed into a dichotomous variable due to the fact that multiple responses were possible.

Table 1: Selected variables for Index creation. Percentage of missing values in the last two columns are rounded to the closest 10. Higher amount of missing corresponds to a deeper red, whereas a lower amount is colored in blue.
* q136 is an multi-response variable that needs to be recoded into multiple dichotomous variables.

		Missing Data in %		
Variable	Explanation	FHE	SUP	Scale
q114_1 [fhe; sup]	Full or part time private med practitioner	70	10	Categorical
q114_2 [fhe; sup]	Full or part-time pharmacist	70	10	Categorical
q114_5 [fhe; sup]	full or part time veterinarian	70	10	Categorical
q116 [fhe; sup]	distance to nearest health center	80	30	Metric
q115 [fhe; sup]	health center in village (alternative to q116)	70	10	Dichotomous
q117 [; sup]	how long to walk to nearest health center (alternative to q116)	100	100	Metric
q121 [fhe; sup]	distance to nearest health post	90	50	Metric
q120 [fhe; sup]	health post in village (alternative to q121)	70	10	Dichotomous
q122 [; sup]	how long to walk to nearest health post (alternative to q121)	100	100	Metric
q126 [fhe; sup]	distance to maternity center	80	30	Metric
q125 [fhe; sup]	maternity center in village (alternative to q126)	70	10	Dichotomous
q127 [; sup]	how long to walk to nearest maternity center (alternative to q126)	100	100	Metric
q130 [fhe; sup]	distance to hospital	80	100	Metric
q129 [fhe; sup]	hospital in village (alternative to q130)	70	100	Dichotomous
q131 [; sup]	how long to walk to nearest hospital (alternative to q130)	100	100	Metric
q134 [fhe; sup]	distance to nearest bus stop	70	100	Metric
q135 [; sup]	how long to walk to nearest bus stop (alternative to q134)	100	100	Metric
q136 [fhe; sup]	drinking water source for most in village	70	10	Categorical*
q137 [fhe; sup]	distance to water	70	10	Metric
q138 [; sup]	how long to walk to water (alternative to q137)	100	90	Metric
q142_1 [fhe; sup]	presence of satellite	70	100	Dichotomous
q142_2 [fhe]	presence of cell phones	70	100	Dichotomous
q143 [fhe; sup]	Electricity	70	100	Dichotomous
q143_2 [fhe; sup]	internet cafe present	70	100	Dichotomous
q192 [fhe; sup]	primary school in village	70	10	Dichotomous

Table 2: Selected variables for Index creation. Percentage of missing values in the last two columns are rounded to the closest 10. Higher amount of missing corresponds to a deeper red, whereas a lower amount is colored in blue.

Variable	Explanation	Missing Data in %
q116	distance to nearest health center	20
q126	distance to maternity center	20
q137	distance to water	10
dist_to_district_center	Distance to district center	0
Distance	Distance to road	0
time_to_school	Distance to School in Minutes.	10
electricity	Self-reported electricity.	10

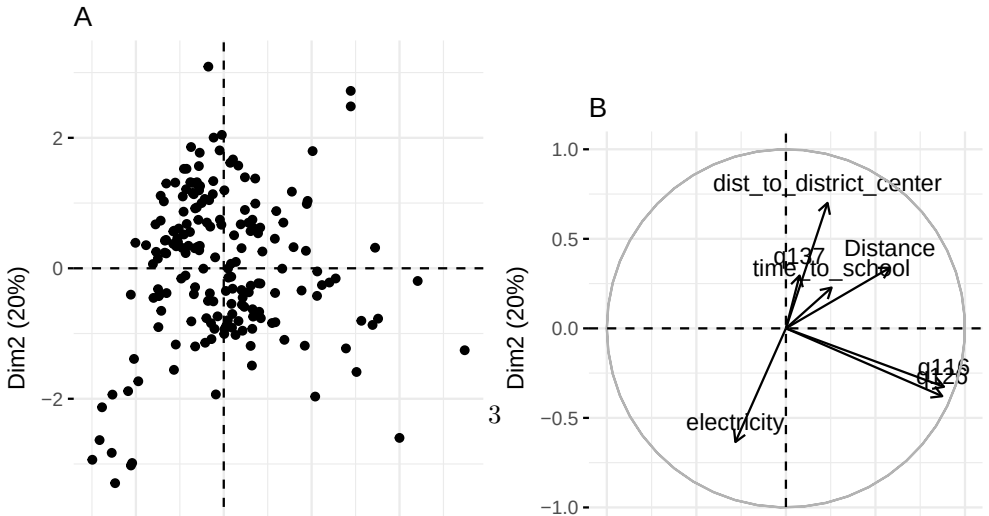
PCA

Table 2 shows the variables used for PCA. Missing values in the SUP data are replaced by FHE if available. If both have non-missing values, the mean is taken between both of them. There are 91 missing values (33%). Further imputation approaches were not used.

Table 3: Eigenvalues of principal components.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.0834219	29.763170	29.76317
comp 2	1.3967043	19.952919	49.71609
comp 3	1.0104872	14.435531	64.15162
comp 4	0.9650090	13.785843	77.93746
comp 5	0.7479918	10.685597	88.62306
comp 6	0.6466129	9.237327	97.86039
comp 7	0.1497729	2.139613	100.00000

Table 3 shows the eigenvalues and percentage of variances of single components. Following the rule of thumb that the eigenvalue needs to be greater than 1 to consider dimensions as meaningful, only the first two components are useful in this context. Figure **??** shows the distribution of individuals (Panel A) and variables (Panel B) along those dimensions. Both graphs indicate two clear dimensions, but its especially Figure **??** which makes clear that q126 and q116 are almost orthogonal to the remaining three variables. The correlation plot highlights this (Figure 2). Hence, it is recommended to retain the two first dimensions. The first dimension measure the distance to the closest health facilities, whereas the second dimension captures the distance to the nearest water extraction and the district center. The distance to the closest road is correlated with both dimensions.



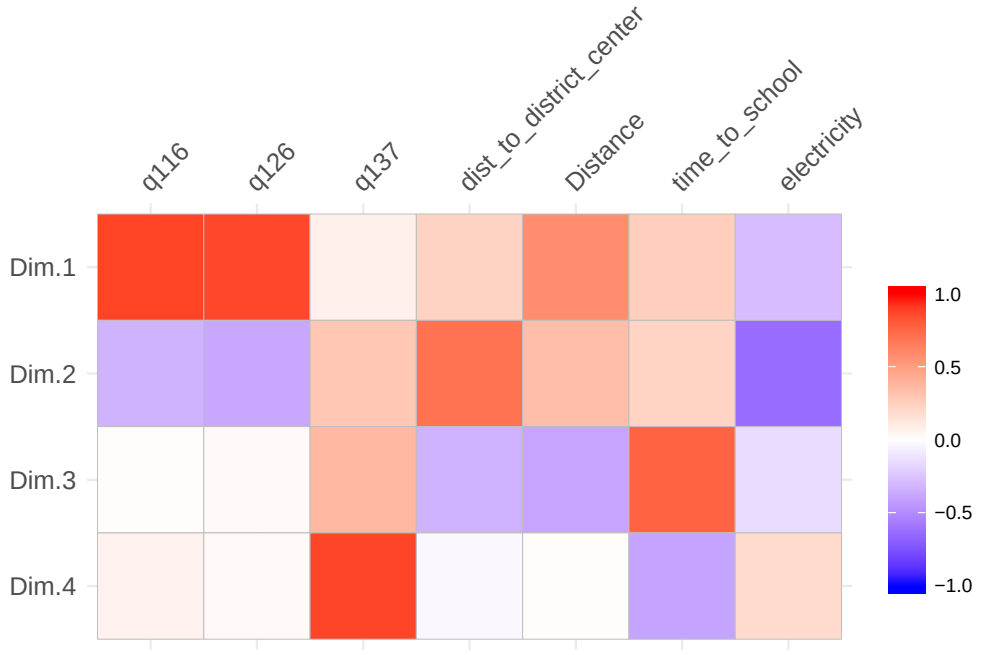


Figure 2: Correlation plot between variables and components.

Table 4: Selected variables for Index creation. Percentage of missing values in the last two columns are rounded to the closest 10. Higher amount of missing corresponds to a deeper red, whereas a lower amount is colored in blue.

Variable	Explanation	Missing
q114_1	Full or part time private med practitioner	10
q114_2	Full or part-time pharmacist	10
q114_5	full or part time veterinarian	10
q115	health center in village (alternative to q116)	10
q120	health post in village (alternative to q121)	10
q125	maternity center in village (alternative to q126)	10
q129	hospital in village (alternative to q130)	70
q136	drinking water source for most in village	10
q142_1	presence of satellite	70
q142_2	presence of cell phones	70
q143	Electricity	70
q192	primary school in village	10

MCA

For the MCA, missing data of SUP is replaced by FHE if available. If both datasets provide non-missing values for a village, only the SUP is considered. Furthermore, missing data and the "Don't Know" category are recoded as extra category to reduce missing data to zero.

Table 5: Eigenvalues MCA.

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.7915379	35.8076653	35.80767
dim 2	0.2240024	10.1334404	45.94111
dim 3	0.1563692	7.0738466	53.01495
dim 4	0.1008094	4.5604255	57.57538
dim 5	0.0922420	4.1728530	61.74823
dim 6	0.0869932	3.9354075	65.68364
dim 7	0.0761190	3.4434793	69.12712
dim 8	0.0670828	3.0346980	72.16182
dim 9	0.0610398	2.7613223	74.92314
dim 10	0.0578269	2.6159780	77.53912
dim 11	0.0532452	2.4087094	79.94783
dim 12	0.0501019	2.2665150	82.21434
dim 13	0.0482737	2.1838118	84.39815
dim 14	0.0443601	2.0067667	86.40492
dim 15	0.0425829	1.9263701	88.33129
dim 16	0.0402232	1.8196199	90.15091
dim 17	0.0343249	1.5527929	91.70370
dim 18	0.0324739	1.4690559	93.17276
dim 19	0.0275925	1.2482331	94.42099
dim 20	0.0263082	1.1901337	95.61112
dim 21	0.0224123	1.0138911	96.62502
dim 22	0.0200807	0.9084123	97.53343
dim 23	0.0182537	0.8257625	98.35919
dim 24	0.0166532	0.7533612	99.11255
dim 25	0.0132642	0.6000483	99.71260
dim 26	0.0025112	0.1136000	99.82620
dim 27	0.0021953	0.0993126	99.92551
dim 28	0.0016466	0.0744875	100.00000

MCA with all variables

```
##        Dim 1      Dim 2
## 5   3.104724 0.05416516
## 36  3.104724 0.05416516
## 56  3.104724 0.05416516
## 60  3.104724 0.05416516
## 65  3.104724 0.05416516
## 71  3.104724 0.05416516
## 74  3.104724 0.05416516
## 86  3.104724 0.05416516
## 99  3.104724 0.05416516
## 122 3.104724 0.05416516
## 123 3.104724 0.05416516
## 125 3.104724 0.05416516
## 139 3.104724 0.05416516
## 147 3.269441 0.67561943
## 156 3.104724 0.05416516
```

```
## 201 3.104724 0.05416516
## 202 3.104724 0.05416516
## 218 3.104724 0.05416516
## 223 3.104724 0.05416516
## 243 3.104724 0.05416516
## 258 3.104724 0.05416516
```

```
##     q114_1 q114_2 q115 q120 q125 q129 q136_cat_2 q136_cat_3 q136_cat_4
## 147     NA     NA   NA   NA   NA   NA         NA         NA         NA
##     q136_cat_5 q136_cat_6 q136_cat_7 q136_cat_8 q136_cat_9 q142_1 q142_2
## 147         NA         NA         NA         NA         NA      0      0
##     q143 q192 q114_5
## 147    0   NA     NA
```

Table 4 shows that several variables have a large percentage of missing values. These missing values have a huge influence on the MCA as Panel B in Figure 3 demonstrates. When only looking at the first dimension, there are 21 extreme outliers on the x axis (dim1 > 3). These are due to the fact, that these variables miss values for (almost) all variables considered here.

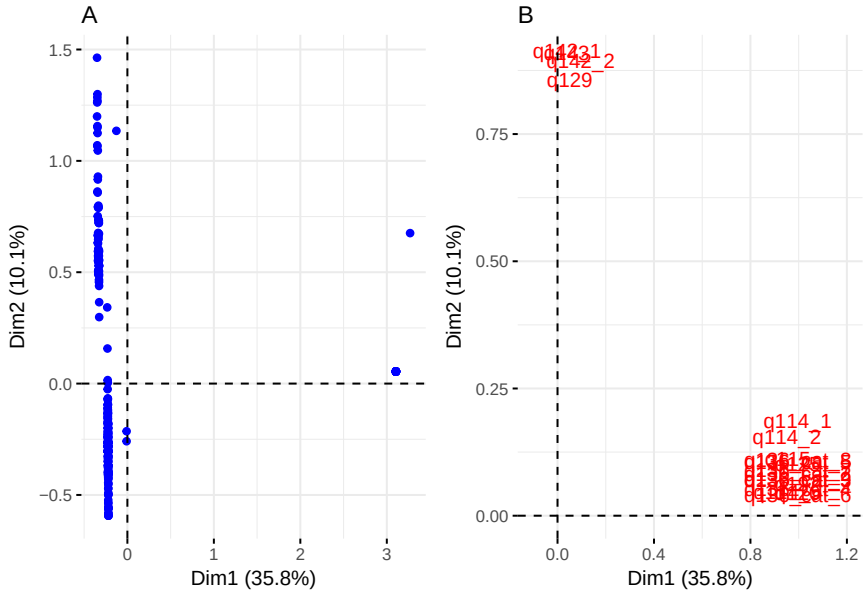


Figure 3: Panel A shows the distribution of individuals and Panel B of variables over two largest components.

MCA excluding variables with a lot of missing values

When excluding the variables with a high amount of missing values and excluding the outliers, the variance captured by the first dimension reduces to 13.5%.

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.1614304	13.452536	13.45254
dim 2	0.1294399	10.786660	24.23920
dim 3	0.1097302	9.144181	33.38338
dim 4	0.0935547	7.796226	41.17960
dim 5	0.0766899	6.390822	47.57042
dim 6	0.0728101	6.067506	53.63793
dim 7	0.0703213	5.860105	59.49804
dim 8	0.0666690	5.555747	65.05378
dim 9	0.0618164	5.151364	70.20515
dim 10	0.0597579	4.979822	75.18497
dim 11	0.0561250	4.677087	79.86206
dim 12	0.0521853	4.348771	84.21083
dim 13	0.0456214	3.801786	88.01261
dim 14	0.0412684	3.439036	91.45165
dim 15	0.0386188	3.218237	94.66989
dim 16	0.0244154	2.034614	96.70450
dim 17	0.0222054	1.850450	98.55495
dim 18	0.0173406	1.445048	100.00000

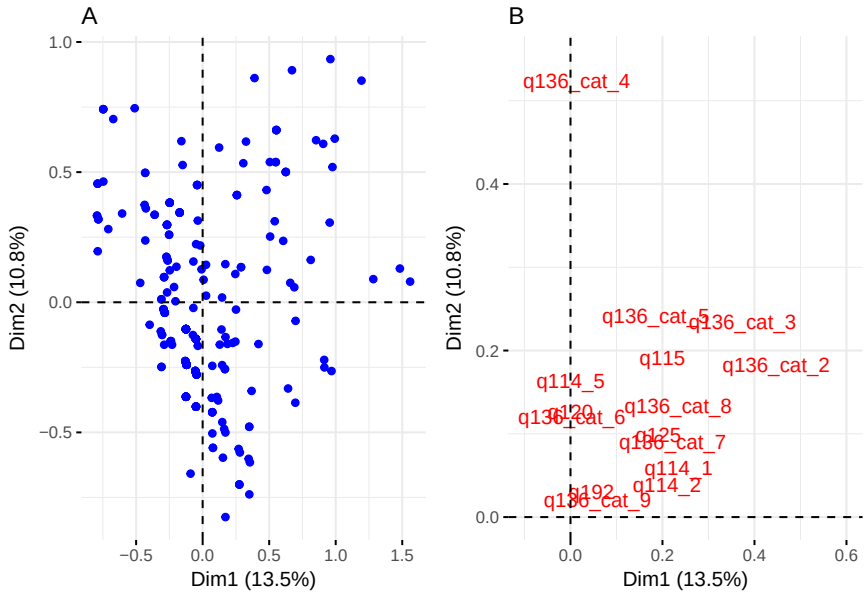


Figure 4: Panel A shows the distribution of individuals and Panel B of variables over two largest components.

Annex

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 16299)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] bindrcpp_0.2.2          raster_2.6-7
##  [3] doParallel_1.0.11       iterators_1.0.9
##  [5] foreach_1.4.4           kableExtra_0.8.0
##  [7] knitr_1.20              gridExtra_2.3
##  [9] factoextra_1.0.5        FactoMineR_1.40
## [11] ggfortify_0.4.4         corrplot_0.84
## [13] rgeos_0.3-26            rgdal_1.2-18
## [15] sp_1.2-7                xlsx_0.5.7
## [17] xlsxjars_0.6.1          rJava_0.9-9
## [19] readstata13_0.9.0       ggcorrplot_0.1.1
## [21] sebstoolkit2_0.0.0.9000 forcats_0.3.0
## [23] stringr_1.3.0           dplyr_0.7.4
## [25] purrr_0.2.4             readr_1.1.1
## [27] tidyr_0.8.0             tibble_1.4.2
## [29] ggplot2_2.2.1           tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.3.1           jsonlite_1.5         viridisLite_0.3.0
##  [4] modelr_0.1.1         assertthat_0.2.0     highr_0.6
##  [7] cellranger_1.1.0     yaml_2.1.18          ggrepel_0.7.0
## [10] pillar_1.2.1         backports_1.1.2      lattice_0.20-35
## [13] glue_1.2.0           digest_0.6.15        RColorBrewer_1.1-2
## [16] rvest_0.3.2          colorspace_1.3-2     htmltools_0.3.6
## [19] plyr_1.8.4           psych_1.8.3.3        pkgconfig_2.0.1
## [22] broom_0.4.4          haven_1.1.1          scales_0.5.0
## [25] ggpubr_0.1.6         lazyeval_0.2.1       cli_1.0.0
## [28] mnormt_1.5-5         magrittr_1.5         crayon_1.3.4
## [31] readxl_1.1.0         evaluate_0.10.1      nlme_3.1-131.1
## [34] MASS_7.3-49          xml2_1.2.0           foreign_0.8-69
## [37] tools_3.4.4          data.table_1.10.4-3  hms_0.4.2
## [40] munsell_0.4.3        cluster_2.0.6        flashClust_1.01-2
## [43] compiler_3.4.4       rlang_0.2.0          grid_3.4.4
## [46] rstudioapi_0.7       leaps_3.0            labeling_0.3
## [49] rmarkdown_1.9        gtable_0.2.0         codetools_0.2-15
```

```
## [52] reshape2_1.4.3       R6_2.2.2             lubridate_1.7.4
## [55] bindr_0.1.1          rprojroot_1.3-2      stringi_1.1.7
## [58] Rcpp_0.12.16         scatterplot3d_0.3-41
```