Predicting the household composition from TV viewing

ETH DAS applied statistics workshop

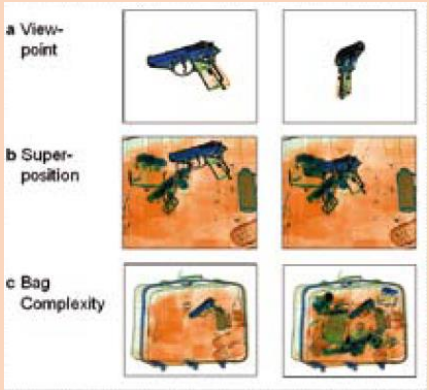
Rafael Lüchinger / 2018-09-10

mediapulse

About Me

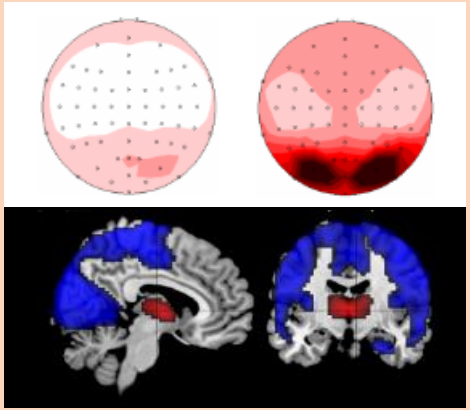
Master Psychology, UZH

Visual Cognition



PhD Neuropsychology, UZH

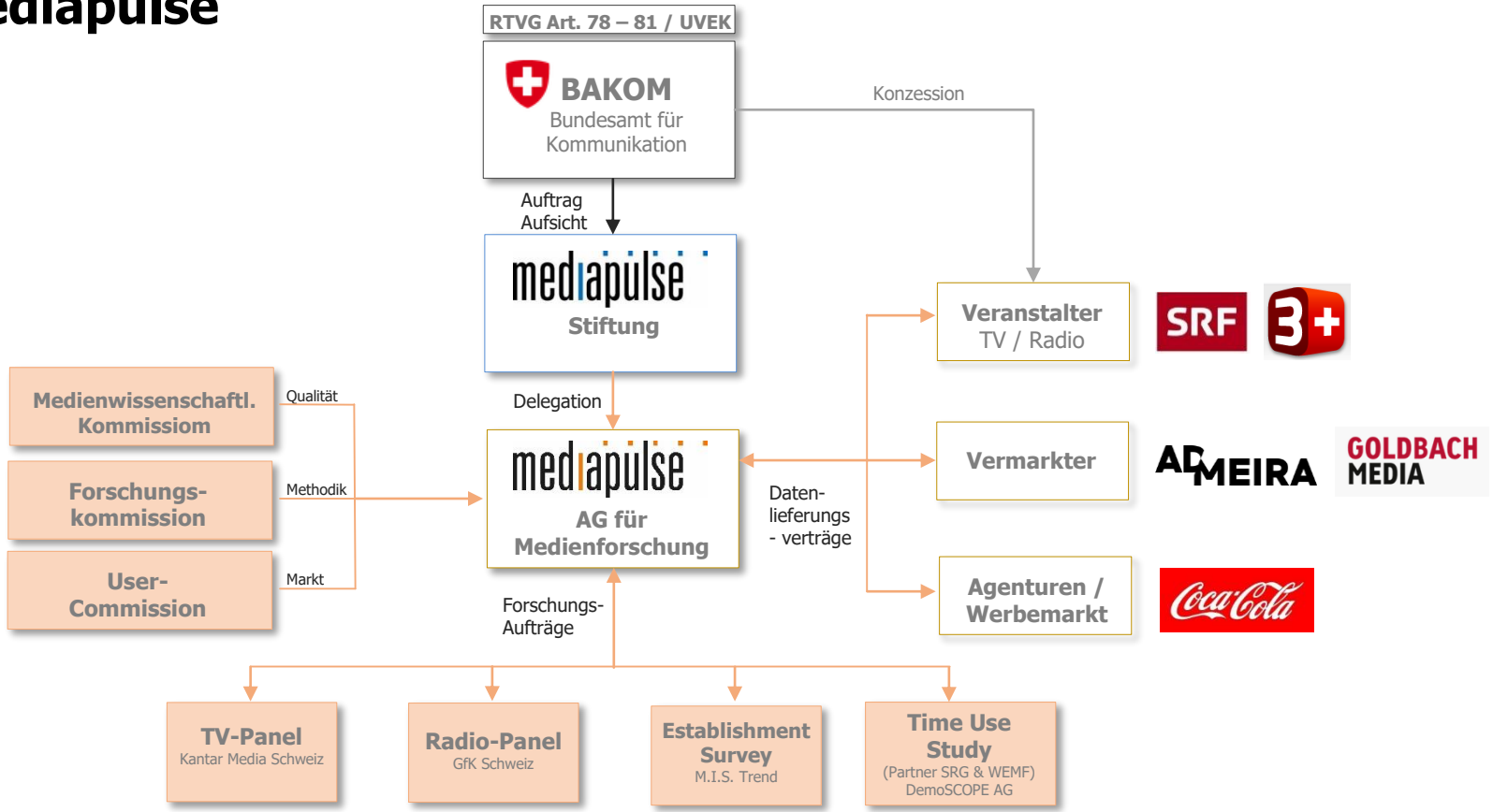
Thalamocortical Interaction



Data Scientist, Mediapulse AG

Radio & TV Nutzung

Mediapulse


RTVG Art. 78 – 81 / UVEK
BAKOM
Bundesamt für Kommunikation
Konzession
Auftrag
Aufsicht
mediapulse
Stiftung
Veranstalter
TV / Radio
SRF
3+
Medienwissenschaftl. Kommissiom
Qualität
Forschungs-kommission
Methodik
User-Commission
Markt
Delegation
mediapulse
AG für Medienforschung
Daten-lieferungs - verträge
Vermarkter
AD MEIRA
GOLDBACH MEDIA
Agenturen / Werbemarkt
Coca-Cola
Forschungs-Aufträge
TV-Panel
Kantar Media Schweiz
Radio-Panel
GfK Schweiz
Establishment Survey
M.I.S. Trend
Time Use Study
(Partner SRG & WEMF)
DemoSCOPE AG

[Die Zahlen gelten für das Jahr 2017]

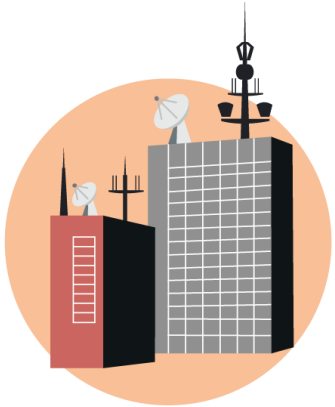
So funktioniert die TV-Nutzungsforschung

Universum
(*3'335'000 TV-HH in CH*)

Stichprobe
(*2008 HH im Panel im Ø*)
(4505 Personen mit G im *im Ø*)

Messung
(*People Meter in HH*)

Fernsehprogramme



Fernsehprogramme zur Referenzierung
(*389 gemessene Programme und 561 registrierte Audio-streams*)

Haushaltsinformationen
Messdaten

Sendeprotokolle

KANTAR MEDIA

Referenzierung
(*2 Aufzeichnungsstationen*)
Audiomatching/Hochrechnung/
Datenkontrollen

mediapulse

Datenkontrollen
Datenveredelung
(*72 Sendeprotokolle*)

Publikation

Kunden

Auswertung der Daten
mit «Instar Analytics»

Die Messtechnologie beruht auf Audiomatching

Das Prinzip des Audiomatching (= Audioabgleich):
Die Tonspur aus dem Panel-Haushalt wird mit den Tonspuren aller referenzierten Sender abgeglichen und zugeordnet.

Im Haushalt wird die Tonspur des Fernseh-konsums registriert

In der Zentrale werden alle Tonspuren der relevanten Sender registriert

Ein Abgleich der beiden Quellen ermöglicht die Zuordnung der Fernsehnutzung im Haushalt

Personen Messung

Panelmitglieder anmelden:
Sobald der Fernseher eingeschaltet wird, erscheint im Display „Wer ist anwesend“. Bitte drücken Sie darauf Ihre Personen-Taste.
Dies gilt für alle anwesenden Personen.

Bevor Sie in die Ferien gehen:
Drücken Sie bitte kurz nach der letzten Nutzung des TV-Gerätes die „Ferien“-Taste. Sobald auf dem Display „Urlaub bestätigen“ erscheint, drücken Sie nochmals die Taste „Ferien“.

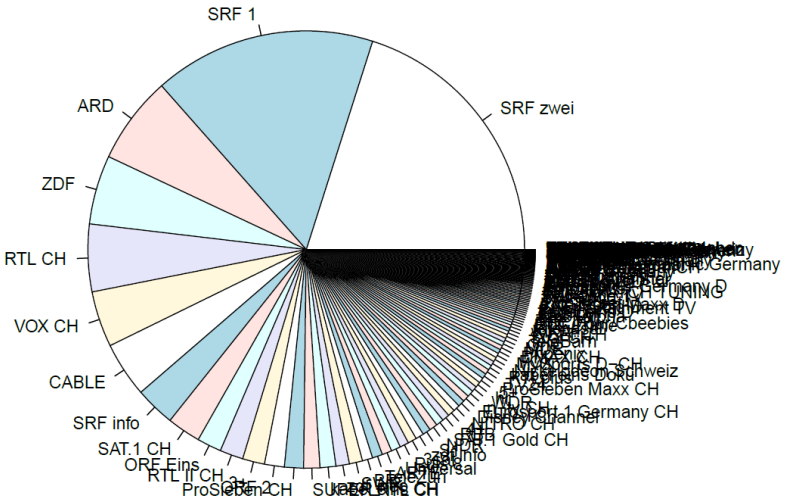
Keine Zuschauer:
Falls das TV-Gerät im Betrieb ist, aber niemand zuschaut, drücken Sie bitte die Taste „Keine Zuschauer“. Dies gilt auch dann, wenn Sie das TV-Gerät zum Radio/CD hören benutzen.

Grenze des Stichprobenansatzes

- Es kommt vor, dass ein Sender zu einer bestimmten Zeit keinen Kontakt aufweist (0-Rating).
- In Wirklichkeit schaut wahrscheinlich irgend jemand das Programm.
- Das TV-Panel ist zu klein um Nutzung mit geringer Reichweite zuverlässig zu erfassen.
- Eine Aufstockung des Panels ist sehr teuer und ineffektiv.
- Das TV Panel ist im internationalen Vergleich relativ gross. Der Schweizer TV Markt ist aber stark fragmentiert.

- Statistische Modellierung
- Anreicherung durch Set-Top-Box Daten (hunderttausende Schweizer Haushalte)

Schweizer TV Markt

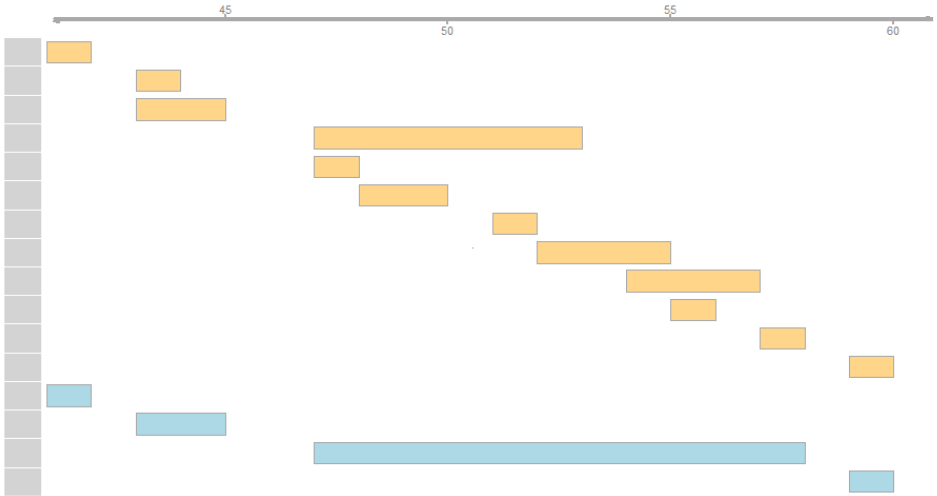


TV Viewing Data

	day	pin	weight	hhsize	chn.name	start	end	dur	act	prg.dur	title	prgtyp	genre
1:	2017-01-01	901	1.162	4	RTS Un	44330	49473	5144	live	5375	Concert du Nouvel An	program	music
2:	2017-01-01	901	1.162	4	RTS Un	49474	49496	23	live	23	2BLS LIVRE JUNGLE 0101	special	trailer
3:	2017-01-01	901	1.162	4	RTS Un	49497	49720	224	live	224	Ensemble	program	info
---													
28220687:	2017-12-31	624001	0.882	1	RTS Un	52879	52919	41	live	1422	Mister Bean (r)	program	show
28220688:	2017-12-31	624001	0.882	1	RTL9 CH	59195	59429	235	live	235	PUB RTL9 SUISSE / 14:55	ad	commercial
28220689:	2017-12-31	624001	0.882	1	RTL9 CH	61921	62169	249	live	249	PUB RTL9 SUISSE / 16:30	ad	commercial

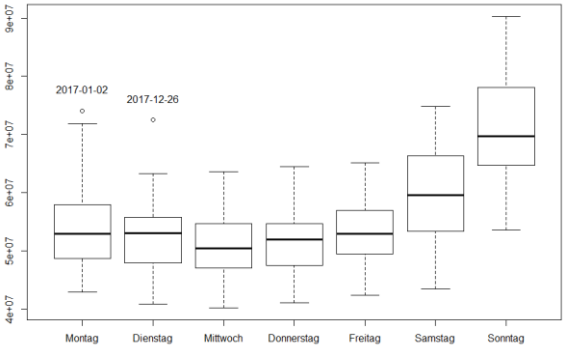
Individual level

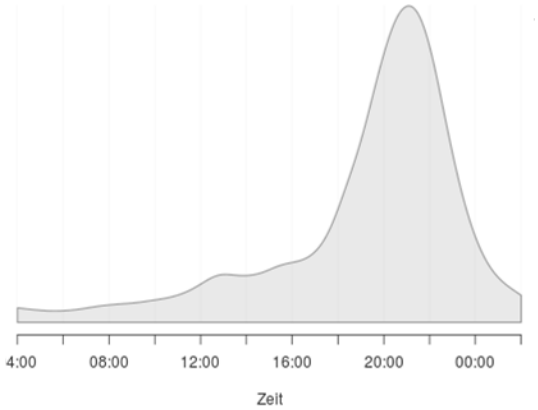
household level



Features Selection

A. Viewing Time





B. Channel Groups

	chn.type	chn.name
1:	Arts	3sat
2:	Arts	ARD-alpha
3:	Arts	ARTE
4:	Arts	Biography Channel
5:	Arts	Classica

461:	Sport	Sportitalia
462:	Sport	Sportitalia 24
463:	Sport	Teleclub Zoom
464:	Sport	Trace Sports
465:	Sport	ULS

C. Program Genre

genre	label	description
0	missing	N/A
1	news	Aktualität
2	info	Magazine, Information, Ratgeber
3	series	Serien
4	talk	Talk-Shows
5	music	Musiksendungen
6	kids	Kinder, Jugend
7	movie	Film
8	show	Unterhaltung, Shows
9	sport	Sport
96	trailer	Promos, Trailer
97	service	Servicesendungen
98	other	Diverse, übrige Sendungen
99	commercial	Werbung

Train & Test Data

```
> head(hh.composition[, -c(2,4:6)])
  hh hhsize age_1 age_2 age_3 age_4 age_5 age_6 age_7 age_8 sex_1 sex_2 sex_3 sex_4 sex_5 sex_6 sex_7 sex_8
1  6      2    70    67     0     0     0     0     0     0     1     2     0     0     0     0     0     0
2  9      4    55    50     0    21    17     0     0     0     1     2     0     1     2     0     0     0
3 14      2    71    72     0     0     0     0     0     0     1     2     0     0     0     0     0     0
4 20      2    59    49     0     0     0     0     0     0     1     2     0     0     0     0     0     0
5 21      2    63    52     0     0     0     0     0     0     1     2     0     0     0     0     0     0
6 35      1    72     0     0     0     0     0     0     0     2     0     0     0     0     0     0     0
```

```
> head(predictors[, 1:8])
  hh day_mofr_02to08 day_mofr_08to12 day_mofr_12to14 day_mofr_14to19 day_mofr_19to23 day_mofr_23to02 day_
1  6    0.0000000000    0.0000000000      0.00000000     0.000493374     0.074948310     0.008649273
2  9    0.0000000000    0.0000000000      0.03518617     0.018886544     0.028364757     0.055750337
3 14    0.0007592294    0.0002186174      0.00000000     0.000000000     0.046034213     0.002020092
4 20    0.0266360558    0.0352911739      0.06770617     0.045673898     0.113524350     0.021275504
5 21    0.0119140911    0.0359914061      0.04071607     0.050380850     0.086132773     0.033098564
6 35    0.0035188132    0.0025159986      0.01183000     0.013437527     0.004811246     0.003528247
```

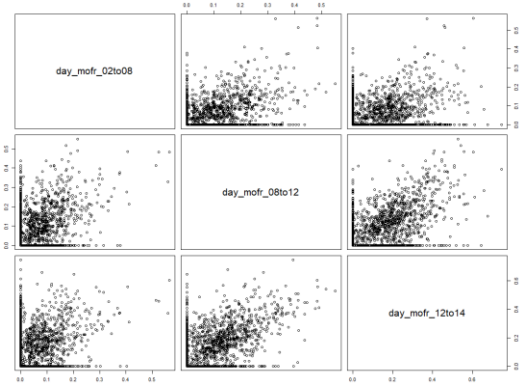
```
> set.seed(1)
> d <- setNames(split(d, runif(nrow(d)) > .6), c("train","test"))
> tbl(d$train$hhsize)
     hhsize1 hhsize2 hhsize3 hhsize4 hhsize5 total
n        430     393     162     166      67  1218
prop      35      32      13      14       6   100
> tbl(d$test$hhsize)
     hhsize1 hhsize2 hhsize3 hhsize4 hhsize5 total
n        265     258     113      99      51   786
prop      34      33      14      13       6   100
```

chance is 1/5

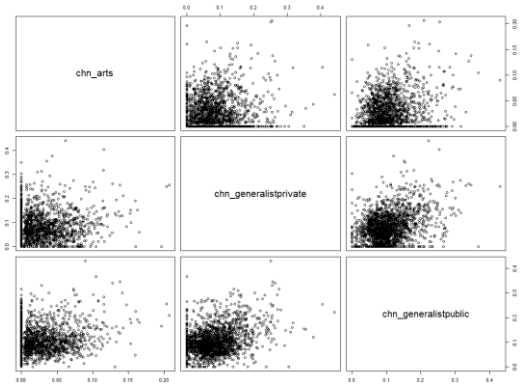
```
> cbind(names(predictors))
      [,1]
 [1,] "hh"
 [2,] "day_mofr_02to08"
 [3,] "day_mofr_08to12"
 [4,] "day_mofr_12to14"
 [5,] "day_mofr_14to19"
 [6,] "day_mofr_19to23"
 [7,] "day_mofr_23to02"
 [8,] "day_saso_02to08"
 [9,] "day_saso_08to12"
[10,] "day_saso_12to14"
[11,] "day_saso_14to19"
[12,] "day_saso_19to23"
[13,] "day_saso_23to02"
[14,] "chn_arts"
[15,] "chn_generalistprivate"
[16,] "chn_generalistpublic"
[17,] "chn_kids"
[18,] "chn_livestileindoor"
[19,] "chn_livestileoutdoor"
[20,] "chn_local"
[21,] "chn_movieseries"
[22,] "chn_music"
[23,] "chn_nature"
[24,] "chn_news"
[25,] "chn_paytv"
[26,] "chn_religion"
[27,] "chn_sport"
[28,] "chn_foreign"
[29,] "chn_swiss"
[30,] "chn_english"
[31,] "chn_french"
[32,] "chn_german"
[33,] "chn_italian"
[34,] "chn_other"
[35,] "prg_commercial"
[36,] "prg_info"
[37,] "prg_kids"
[38,] "prg_missing"
[39,] "prg_movie"
[40,] "prg_music"
[41,] "prg_news"
[42,] "prg_other"
[43,] "prg_series"
[44,] "prg_service"
[45,] "prg_show"
[46,] "prg_sport"
[47,] "prg_talk"
[48,] "prg_trailer"
```

Between features

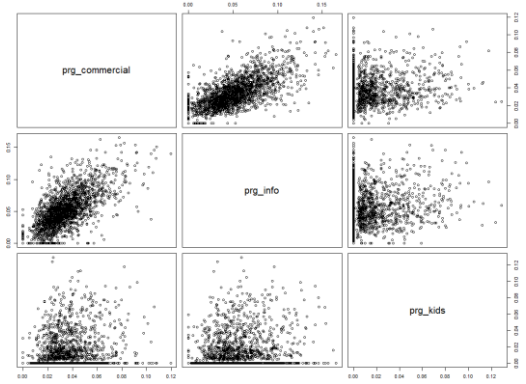
A. Viewing Time

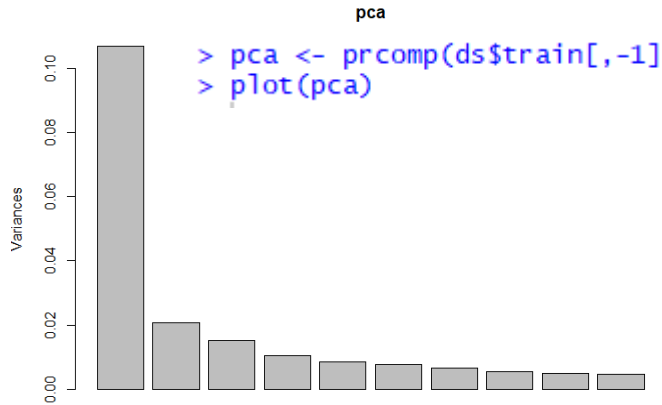


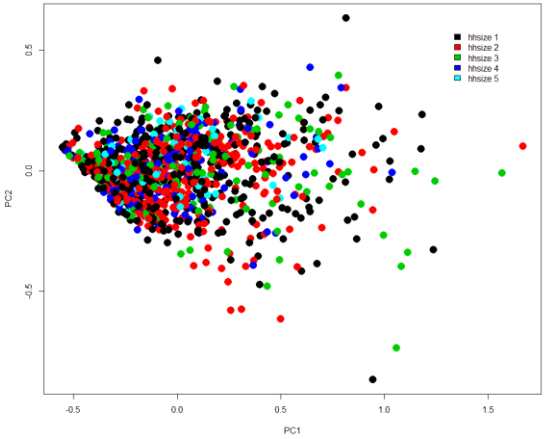
B. Channel Groups



C. Program Genre



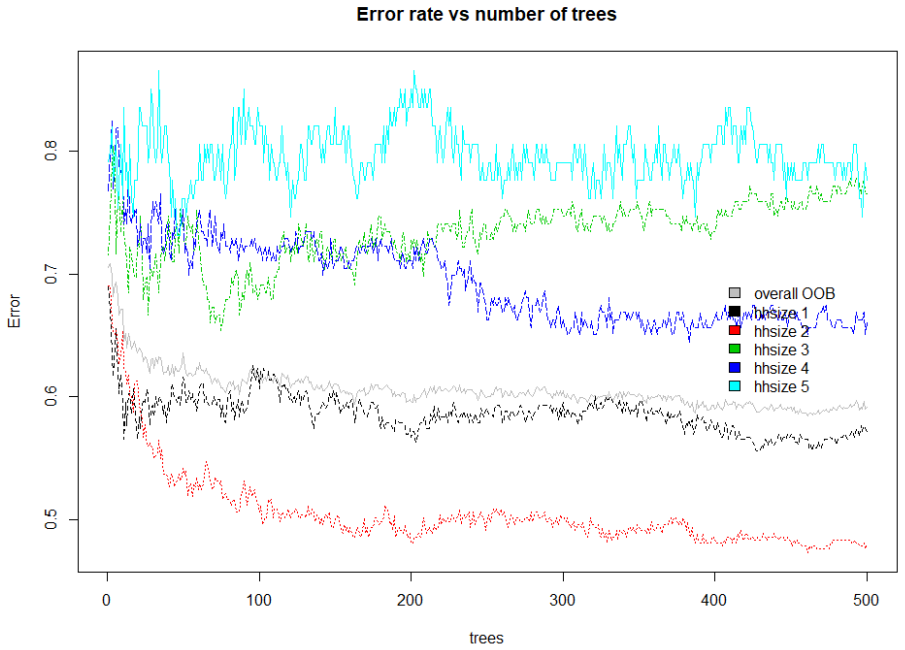




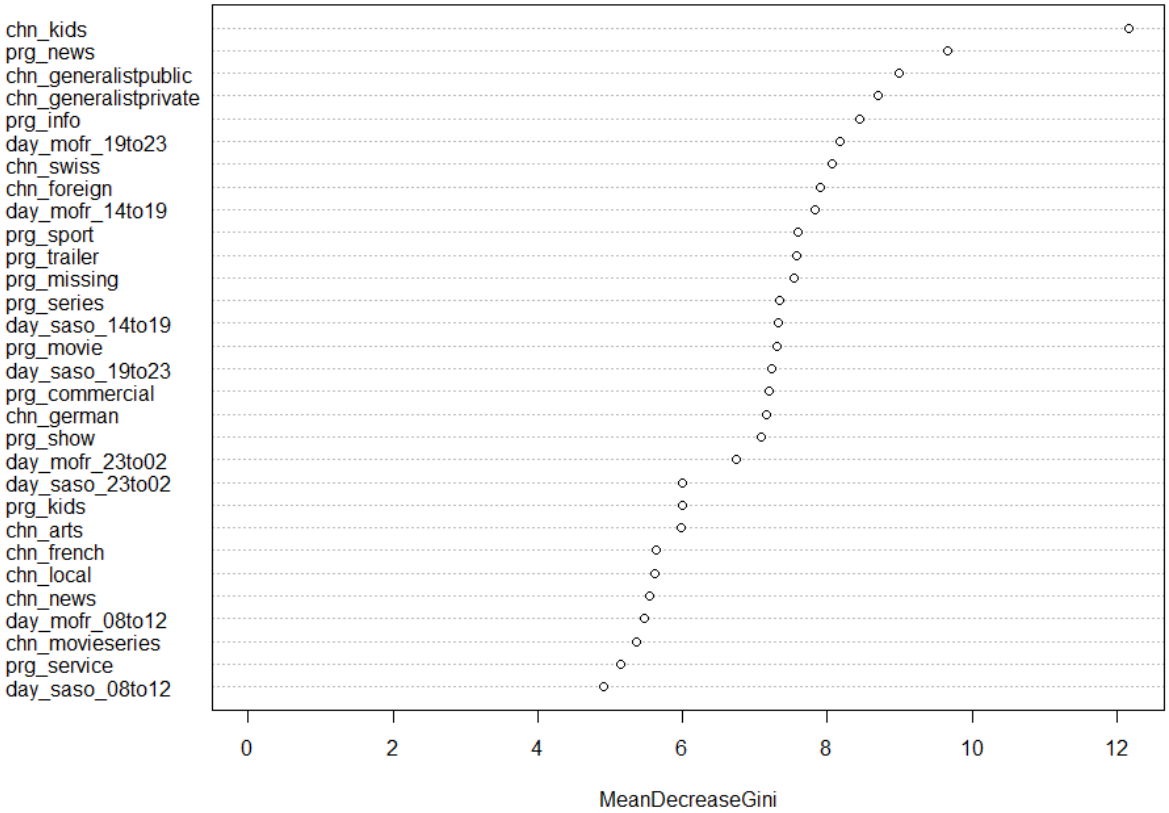
Random Forest

```
> rf <- randomForest(
+   hhsize ~ ., data = ds$train, imortance = TRUE,
+   strata = ds$train$hhsize, sampsize = rep(min(table(ds$train$hhsize)), 5) # 67
+   )
```

```
> performance.rf
$`train`
$`train`$`accuracy`
[1] 0.408046

$`train`$confusion
        hhsize1 hhsize2 hhsize3 hhsize4 hhsize5 class.error
hhsize1     184     133      62      41      10   0.5720930
hhsize2     111     204      36      40       2   0.4809160
hhsize3      31      29      38      40      24   0.7654321
hhsize4      23      30      26      56      31   0.6626506
hhsize5       6      10       8      28      15   0.7761194


$test
$test$`accuracy`
[1] 0.4122137

$test$confusion
          predict
true       hhsize1 hhsize2 hhsize3 hhsize4 hhsize5
  hhsize1      111      81      47      17       9
  hhsize2       57     139      32      20      10
  hhsize3       12      33      31      29       8
  hhsize4       15      22      15      32      15
  hhsize5        3       5       7      24      12
```



Variable Importance



Outlook

- More & better features, based on longer period, e.g. year
- Other Classifier (Multinomial Logistic Model, Linear Discriminant Analysis, Stochastic Gradient Boosting, Support Vector Machine)
- Could we use priors for hh composition by region based on FSO data
- From household size to individual level with age / sex
- Change household size for household composition

- Family (F): a household that consists of two adults, irrespective of their gender.
- Family with children (FC): a household that consists of two adults, irrespective of their gender, with at least one child[2].
- Household (H): a household that consists of more than two adults.
- Household with children (HC): a household that consists of more than two adults with at least one child.
- Single female (SF): a household with only one adult female.
- Single male (SM): a household with only one adult male.
- Single female parent (SPF[3]): a household with only one adult female with at least one child.
- Single male parent (SPM[4]): a household with only one adult male with at least one child.

Pre Family:	Head of Household aged <45, No Children 0-15 in Home
Young Family:	Household contains Children 0-3
Older Family:	Household contains Children 4-15 but none 0-3
Post Family:	Head of Household aged 45+, No Children 0-15 in Home, at least one Household Member working full/part time
Inactive:	Head of Household aged 45+, No Children 0-15 in Home, no Household Member working full/part time

Man, 2016, Statistical Learning of Household Composition from Television Viewing Behaviour

RSNB 2014, Panel Control Review 2014 for BARB

mediapulse